COGNTIVE 2026

The Eighteenth International Conference on Advanced Cognitive Technologies and Applications

ISBN: 978-1-68558-375-0

April 19 - 23, 2026

Lisbon, Portugal

COGNITIVE 2026 Editors

Carlos Peña, Jacobs Institute - Buffalo, USA
Kristina Schaaff, IU International University of Applied Sciences, Germany

COGNITIVE 2026

Forward

The Eighteenth International Conference on Advanced Cognitive Technologies and Applications (COGNITIVE 2026), held on April 19 – 23, 2026, targeted advanced concepts, solutions and applications of artificial intelligence, knowledge processing, agents, as key-players, and autonomy as manifestation of self-organized entities and systems. The advances in applying ontology and semantics concepts, web-oriented agents, ambient intelligence, and coordination between autonomous entities led to different solutions on knowledge discovery, learning, and social solutions.

The conference had the following tracks:

- Brain information processing and informatics
- Artificial intelligence and cognition
- Agent-based adaptive systems
- Applications
- Autonomous systems and autonomy-oriented computing
- Hot topics on cognitive science

Similar to the previous edition, this event attracted excellent contributions and active participation from all over the world. We were very pleased to receive top quality contributions.

We take here the opportunity to warmly thank all the members of the COGNITIVE 2026 technical program committee, as well as the numerous reviewers. The creation of such a high quality conference program would not have been possible without their involvement. We also kindly thank all the authors that dedicated much of their time and effort to contribute to COGNITIVE 2026. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

Also, this event could not have been a reality without the support of many individuals, organizations and sponsors. We also gratefully thank the members of the COGNITIVE 2026 organizing committee for their help in handling the logistics and for their work that made this professional meeting a success.

We hope COGNITIVE 2026 was a successful international forum for the exchange of ideas and results between academia and industry that will promote further progress in the area of cognitive technologies and applications. We also hope that Lisbon provided a pleasant environment during the conference and everyone saved some time to enjoy this beautiful city.

COGNITIVE 2026 General Chair

Jaime Lloret Mauri, Universitat Politecnica de Valencia, Spain

COGNITIVE 2026 Steering Committee

Jayfus Tucker Doswell, The Juxtopia Group, Inc., USA
Thomas Ågotnes, University of Bergen, Norway
Muneo Kitajima, Nagaoka University of Technology (Emeritus), Japan
Kristina Schaaff, IU International University of Applied Sciences, Germany

COGNITIVE 2026 Publicity Chair

Francisco Javier Díaz Blasco, Universitat Politècnica de València, Spain
Ali Ahmad, Universitat Politècnica de València, Spain
José Miguel Jiménez, Universitat Politècnica de València, Spain
Sandra Viciano Tudela, Universitat Politècnica de València, Spain

COGNITIVE 2026

Committee

COGNITIVE 2026 General Chair

Jaime Lloret Mauri, Universitat Politecnica de Valencia, Spain

COGNITIVE 2026 Steering Committee

Jayfus Tucker Doswell, The Juxtopia Group, Inc., USA
Thomas Ågotnes, University of Bergen, Norway
Muneo Kitajima, Nagaoka University of Technology (Emeritus), Japan

COGNITIVE 2026 Publicity Chair

Francisco Javier Díaz Blasco, Universitat Politècnica de València, Spain
Ali Ahmad, Universitat Politècnica de València, Spain
José Miguel Jiménez, Universitat Politècnica de València, Spain
Sandra Viciano Tudela, Universitat Politècnica de València, Spain

COGNITIVE 2026 Technical Program Committee

Witold Abramowicz, University of Economics and Business, Poland
Thomas Agotnes, University of Bergen, Norway
Vered Aharonson, University of the Witwatersrand, Johannesburg, South Africa
Thangarajah Akilan, Lakehead University, Canada
Maksim Alekseevich Eremeev, Genentech, USA
Piotr Artiemjew, University of Warmia and Masuria in Olsztyn, Poland
Divya B, SSNCE, India
Thierry Bellet, University Gustave Eiffel, France
Petr Berka, University of Economics, Prague, Czech Republic
Ateet Bhalla, Independent Consultant, India
Mahdi Bidar, University of Regina, Canada
Guy Andre Boy, CentraleSupélec, LGI, Paris Saclay University / ESTIA Institute of Technology, France
Dilyana Budakova, Technical University of Sofia - Branch Plovdiv, Bulgaria
Valerie Camps, Paul Sabatier University - IRIT, Toulouse, France
Yaser Chaaban, Leibniz University of Hanover, Germany
Olga Chernavskaya, P. N. Lebeved Physical Institute, Moscow, Russia
Helder Coelho, Universidade de Lisboa, Portugal
Igor Val Danilov, Academic Center for Coherent Intelligence, Latvia
Angel P. del Pobil, Jaume I University, Spain
Soumyabrata Dev, University College Dublin, Ireland
Jerome Dinet, University of Lorraine, France

Razieh Saremi, Stevens Institute of Technology, USA
Kristina Schaaff, IU International University of Applied Sciences, Germany [emotional computation]
Ljiljana Šerić, University of Split, Croatia
Naavya Shetty, University of Illinois Urbana-Champaign, USA
Paul Smart, University of Southampton, UK
Stanimir Stoyanov, Plovdiv University "Paisii Hilendarski", Bulgaria
Nasseh Tabrizi, East Carolina University, USA
Tiago Thompsen Primo, Samsung Research Institute, Brazil
Gary Ushaw, Newcastle University, UK
Jaap van den Herik, Leiden Centre of Data Science (LCDS) | Leiden University, Leiden, The Netherlands
S.Vidhusha, SSN College of Engineering, Chennai, India
Emilio Vivancos, Valencian Research Institute for Artificial Intelligence (VRAIN) | Universitat Politecnica de Valencia, Spain
Han Wang, Beijing Institute of Technology Zhuhai / Zhuhai Institute of Advanced Technology - Chinese Academy of Sciences, China
Xianzhi Wang, University of Technology Sydney, Australia
Yingxu Wang, University of Calgary, Canada
Ye Yang, Stevens Institute of Technology, USA
Sule Yildirim Yayilgan, NTNU, Norway
Besma Zeddini, EISTI, France

Copyright Information

For your reference, this is the text governing the copyright release for material published by IARIA.

The copyright release is a transfer of publication rights, which allows IARIA and its partners to drive the dissemination of the published material. This allows IARIA to give articles increased visibility via distribution, inclusion in libraries, and arrangements for submission to indexes.

I, the undersigned, declare that the article is original, and that I represent the authors of this article in the copyright release matters. If this work has been done as work-for-hire, I have obtained all necessary clearances to execute a copyright release. I hereby irrevocably transfer exclusive copyright for this material to IARIA. I give IARIA permission or reproduce the work in any media format such as, but not limited to, print, digital, or electronic. I give IARIA permission to distribute the materials without restriction to any institutions or individuals. I give IARIA permission to submit the work for inclusion in article repositories as IARIA sees fit.

I, the undersigned, declare that to the best of my knowledge, the article is does not contain libelous or otherwise unlawful contents or invading the right of privacy or infringing on a proprietary right.

Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article.

IARIA grants royalty-free permission to the authors to disseminate the work, under the above provisions, for any academic, commercial, or industrial use. IARIA grants royalty-free permission to any individuals or institutions to make the article available electronically, online, or in print.

IARIA acknowledges that rights to any algorithm, process, procedure, apparatus, or articles of manufacture remain with the authors and their employers.

I, the undersigned, understand that IARIA will not be liable, in contract, tort (including, without limitation, negligence), pre-contract or other representations (other than fraudulent misrepresentations) or otherwise in connection with the publication of my work.

Exception to the above is made for work-for-hire performed while employed by the government. In that case, copyright to the material remains with the said government. The rightful owners (authors and government entity) grant unlimited and unrestricted permission to IARIA, IARIA's contractors, and IARIA's partners to further distribute the work.

Table of Contents

The Influence of Extraversion and Neuroticism on Technology Acceptance and Cognitive Dissonance in LLM Usage

Alicia Unland, Marc-André Heidelmann, Kristina Schaaff
IU International University of Applied Sciences, Erfurt, Germany
e-mail: {marc-andre.heidelmann | kristina.schaaff}@iu.org

***Abstract*—In our cross-sectional study, we examine how personality traits influence students' acceptance of Large Language Models (LLMs) in higher education. Building on the Technology Acceptance Model, the Big Five framework, and Cognitive Dissonance Theory, we focus on Extraversion and Neuroticism as predictors of LLM Usage. We conducted an online survey with 120 psychology students, measuring personality traits, Technology Acceptance, LLM Usage, and Cognitive Dissonance. Extraversion showed no significant association with either acceptance or dissonance. Neuroticism had a negative direct effect on Technology Acceptance, and we observed a small positive indirect effect via Cognitive Dissonance that warrants cautious interpretation. These results may help explain why some more neurotic students nonetheless accept LLMs.**

Keywords-technology acceptance; cognitive dissonance; personality traits; extraversion; neuroticism; human–computer interaction.

I. Introduction

Large Language Models (LLMs), such as ChatGPT, Gemini, or Perplexity AI have rapidly entered higher education, transforming how people access, generate, and interact with information [1]–[3]. Universities increasingly integrate these tools to support teaching and learning (e.g., [4]–[6]). At the same time, limited transparency and reliability raise concerns about trust, critical evaluation, and responsible use [7]. These tensions raise a central question for both educational practice and research in Human-Computer Interaction (HCI): When and why do students adopt LLMs, and how do personality traits influence this adoption?

Technology Acceptance Models provide well-established frameworks for explaining adoption, emphasizing perceived usefulness, ease of use, and user experience as key predictors [8][9]. However, these models often treat users as homogeneous and overlook how individual differences shape acceptance. Students bring diverse expectations, practices, and psychological traits that influence how they interact with AI technologies.

Prior work shows that personality traits—especially Extraversion and Neuroticism—are associated with trust and technology use, with Extraversion linked to higher acceptance and Neuroticism to skepticism or lower usability perceptions [10]–[12]. [13] found that higher neuroticism and extraversion were associated with a more serious perceived AI-related threat. However, empirical research on personality in the context of LLM use is scarce [14]. The role of Cognitive Dissonance—the discomfort when expectations diverge from experience [15]—also remains underexplored, even though the error-prone and sometimes contradictory nature of LLMs may elicit dissonance that shapes acceptance and continued use. Therefore, understanding how students adopt LLMs requires an interdisciplinary perspective that integrates educational technology, personality psychology, and human–computer interaction. Prior research highlights both the transformative potential of AI in higher education [1][5][16] and challenges around transparency, trust, and responsible use [17][18], underscoring the relevance of personality-driven differences in learners' engagement with LLM-based tools.

In our study, we address these gaps by examining how Extraversion and Neuroticism relate to students' acceptance of LLMs and their experience of Cognitive Dissonance. Drawing on the Technology Acceptance Model [8], the Big Five personality framework [19], and Cognitive Dissonance Theory [15], we model Cognitive Dissonance as a mediator between personality and LLM acceptance and quantify its contribution to the Technology Acceptance of LLMs.

Our study offers three contributions:

1) **Empirical**: we provide quantitative evidence on the role of personality and Cognitive Dissonance in LLM adoption;
2) **Theoretical**: we extend acceptance models with psychological factors to better explain diverse user experiences;
3) **Practical**: we derive implications for designing adaptive, human-centered AI tools that support trust, equity, and well-being in higher education.

The remainder of this paper is structured as follows. In Section 2, we review the theoretical background on AI in education, personality psychology, technology acceptance models, and cognitive dissonance. In Section 3, we present the study design, including research questions, hypotheses, sample characteristics, instruments, and analytical procedures. In Section 4, we report the empirical results of the correlation and mediation analyses. In Section 5, we discuss the findings in relation to existing literature, outline implications for human–computer interaction and educational practice, and address limitations. Finally, in Section 6, we conclude with a summary of key insights and directions for future research.

II. Background

To frame our study, this section reviews three key areas: the integration of AI and LLMs in education (Section II-A), personality psychology with a focus on the personality traits Extraversion and Neuroticism (Section II-B), and established technology acceptance models and their extensions (Section II-C). Moreover, we address Cognitive Dissonance Theory to

 ISBN: 978-1-68558-375-0

explain how discrepancies between expectations and experiences may influence user acceptance in Section II-D.

A. AI in Education

Recent LLMs, such as ChatGPT, Gemini, and Perplexity AI, enable human-like interaction and have been rapidly integrated into both everyday life and academic contexts [2][3], making them powerful tools for education and research. However, concerns remain regarding reliability, transparency, and critical evaluation of generated content [7]. The use of LLMs in higher education is a relatively new and emerging field. Students frequently use text-generating AI tools, such as ChatGPT, primarily because of their ease of access rather than systematic pedagogical integration [7]. Current studies suggest that acceptance of these technologies can be explained through established models of technology adoption, such as the Technology Acceptance Model (TAM) [7], which emphasize user perceptions that significantly shape adoption decisions [20].

B. Personality Psychology and the Big Five Model

Personality refers to systematic interindividual differences in cognition, affect, and behavior [21]. One of the most established frameworks in personality psychology is the Big Five model (OCEAN) [19], which conceptualizes personality along five continuous dimensions: Openness to Experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism [22].

At the intersection of psychology and HCI, personality traits have been increasingly studied as predictors of Technology Acceptance. In our study, we focus on Extraversion and Neuroticism, which are strongly linked to trust, attitudes, and acceptance of AI technologies [10]. Extraversion has been linked to higher acceptance of digital tools and positive evaluations of social technologies [11][23], whereas Neuroticism relates to lower perceived usefulness, reduced trust, and higher uncertainty and skepticism, potentially reducing acceptance [10][24]. From the perspective of Cognitive Dissonance Theory, individuals high in Neuroticism are more prone to negative affect and thus more likely to experience dissonance when system outputs conflict with expectations. In contrast, extraverted individuals typically show more positive affect and resilience in coping with such discrepancies [15][25].

C. Technology Acceptance Models in the Context of LLMs

The TAM, originally proposed by Davis [8], provides a framework for understanding why individuals adopt or reject specific technologies. It highlights two central constructs: *Perceived Usefulness* (the belief that using a technology enhances performance) and *Perceived Ease of Use* (the degree to which a technology is considered effortless to use), which shape user attitudes, behavioral intentions, and actual system usage. Perceived usefulness is generally the strongest direct predictor of Technology Acceptance, while perceived ease of use additionally influences perceived usefulness [8].

TAM has been extended to account for external factors. TAM2 [9] adds social influences (e.g., subjective norms, image) and cognitive processes (e.g., job relevance, output quality, voluntariness) to improve predictive validity. Unified Theory of Acceptance and Use of Technology (UTAUT) [26] integrates multiple frameworks into a comprehensive model. The Media and Technology Usage and Attitudes Scale (MTUAS) [27] broadens the focus by incorporating psychological aspects of everyday technology use, including trust, compatibility, and satisfaction. In parallel, the Digital Technology Acceptance Scale (DTAS) [28], developed as an extension of TAM [9][26], is used to study AI-based systems like LLMs.

While TAM and its extensions have been validated across various technologies, their application to AI-based systems, particularly LLMs, raises new questions. Classical constructs, such as perceived usefulness and perceived ease of use, remain relevant, but additional factors come into play: transparency, trust, and ethical considerations are central to AI acceptance [7], and the black-box nature of LLMs can undermine trust, as users may struggle to evaluate the reliability of outputs. Studies indicate that social influences and prior experience shape acceptance of generative AI in higher education [20]. Students often adopt LLMs for pragmatic reasons such as efficiency and accessibility. Yet, their long-term integration depends on whether the tools are perceived as credible and aligned with academic values. Furthermore, acceptance is influenced by personality traits (e.g., Extraversion, Neuroticism) and cognitive processes (e.g., dissonance when outputs conflict with expectations), which are not fully captured by classical TAM variables. Therefore, in educational settings, extended models, such as DTAS, are especially relevant, as they incorporate behavioral intention and enjoyment. This is critical for understanding whether students use LLMs, why they integrate them into their study practices, and how they experience them, since preferences may vary across different learner profiles [28].

D. Cognitive Dissonance in LLM Usage

Cognitive Dissonance Theory [15] describes the psychological discomfort that arises when individuals hold conflicting cognitions (e.g., expectations vs. actual experiences), motivating efforts to restore consistency by changing attitudes, behaviors, or avoiding dissonant situations.

Applied to LLM Usage, this framework highlights new challenges when performance expectations are not met: the inconsistency between expectations and outcomes can trigger negative emotions [25]. In the context of students using LLMs, dissonance may arise when expectations about reliability or accuracy do not align with received outputs [29]. Since LLMs are prone to generating errors or false information, confronting misinformation can provoke dissonance [30], especially when students rely heavily on the correctness of generated content but later realize limitations, undermining both trust in the system and self-confidence. Evaluating and correcting such misinformation is central for individual well-being and broader reliability of technology in educational settings [31]. Festinger's principles thus provide a powerful lens on emotional and cognitive reactions in interaction with LLMs, especially in education, where unresolved dissonance can shape learning

behaviors, attitudes toward technology, and long-term patterns of adoption or avoidance.

III. Study Overview and Design

In our study, we aim to find out how personality traits and Cognitive Dissonance influence the Technology Acceptance of psychology students. After presenting our research questions, we describe the research methodology we applied in our study.

A. Research Questions and Hypotheses

The theoretical background on Technology Acceptance and Cognitive Dissonance suggests that personality traits may shape how students perceive and adopt LLMs. While prior work emphasizes the central role of perceived usefulness and ease of use [8][9], there is limited empirical evidence on how personality-driven factors, such as Extraversion and Neuroticism, influence acceptance in educational AI contexts. Moreover, Cognitive Dissonance may act as an additional explanatory mechanism when expectations and experiences with LLMs diverge [15][25].

Based on the theoretical framework outlined above, Table I summarizes the research questions and corresponding hypotheses examined in this study.

TABLE I. Research questions and hypotheses.

ID	Description
RQ1	How does Extraversion relate to students' acceptance of LLMs?
RQ2	How does Neuroticism relate to Cognitive Dissonance when interacting with LLMs?
RQ3	Does Cognitive Dissonance mediate the relationship between Neuroticism and Technology Acceptance?
RQ4	How do personality traits relate to actual LLM Usage?
H1	Extraversion is positively associated with Technology Acceptance of LLMs.
H2	Neuroticism is positively associated with Cognitive Dissonance when using LLMs.
H3a	Cognitive Dissonance negatively mediates the relationship between Neuroticism and Technology Acceptance.
H3b	(exploratory) Technology Acceptance positively mediates the relationship between Extraversion and LLM Usage.
H4	(exploratory): Neuroticism is negatively associated with Technology Acceptance of LLMs.
H5	(exploratory): Extraversion is negatively associated with Cognitive Dissonance when using LLMs.

Figure 1 presents the direct hypotheses, while Figure 2 shows the mediation hypotheses, integrating personality traits, Technology Acceptance, and Cognitive Dissonance.

B. Data Collection and Participants

Data were collected between May and June 2025 using an online questionnaire. We recruited 127 participants, mostly students from IU International University of Applied Sciences but also from psychology-related social media channels, such as WhatsApp groups. Before participation, all students were informed about the purpose of the study and data protection regulations in accordance with the GDPR, and they provided

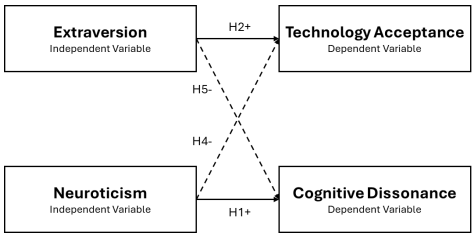


Figure 1. Direct hypotheses. Dashed arrows indicate exploratory hypotheses.

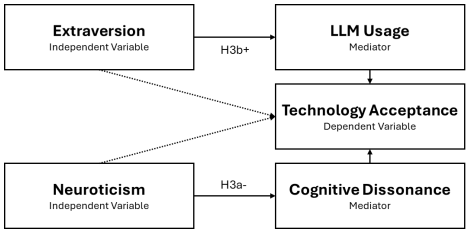


Figure 2. Mediation hypotheses. Solid arrows represent indirect paths, dotted arrows represent direct paths.

informed consent. Participants without a background in psychology and those younger than 18 years were excluded.

We focused on psychology students as they form a particularly relevant population for examining the relationship between personality traits and LLM Usage. Personality constructs, such as the Big Five model, are central to their field of study, which enhances familiarity with self-report instruments and therefore increases data quality. Moreover, psychology students frequently engage in text-based academic tasks (e.g., essays, literature reviews) and are therefore likely to experiment with tools like ChatGPT. Using a relatively homogeneous sample also reduces potential confounding effects of disciplinary differences and allows a clearer examination of personality-driven influences.

We used a quantitative, correlational design with a cross-sectional online survey, where participants rated items on Likert-type scales. Experimental manipulation was not used, as personality traits cannot be manipulated in this context. Therefore, a correlational design was adopted to examine the relationships between personality traits, Technology Acceptance, Cognitive Dissonance, and LLM Usage. An a-priori power analysis indicated a minimum sample size of $N = 84$ to test **H1** and **H2** (medium effects: $\rho = .30$, $\alpha = .05$, $1 - \beta = .80$). For the mediation hypotheses **H3a** and **H3b**, assuming a small-to-medium effect size ($f^2 = 0.06$), a minimum sample size of $N = 120$ was required to achieve adequate power ($\alpha = .05$, $1 - \beta = .80$). Since all hypotheses were tested within the same study, a total sample size of $N = 120$ participants was necessary.

C. Data Preparation and Cleaning

All items were measured on Likert-type or frequency scales. For each construct, we computed mean scores across the respective items and treated these composite scores as approximately interval-scaled. We inspected missing values and removed cases with excessive missing data listwise; scale

scores were computed, handling remaining missing responses on single items if at least half of the items for a given construct were available. Internal consistencies (Cronbach's α) were calculated to ensure reliability of the aggregated scales.

To ensure data quality, we embedded two attention checks in the questionnaire and included only participants who answered both correctly. In addition, we applied a lower time threshold of two minutes. Cases with substantially shorter times were checked for plausibility and excluded if necessary.

The Usage Frequency Scale [27], originally with ten response categories, was reduced to nine categories in the present study to ensure compatibility with the other scales.

D. Sample Characteristics

After data preparation and cleaning, the final sample included $N = 120$ valid cases for analysis. Table II summarizes the demographic data of the study participants. Participants were between 18 and 45 years old, with a mean age of 24.91 years (SD = 5.48).

TABLE II. SOCIODEMOGRAPHIC DATA OF THE FINAL SAMPLE ($n = 120$).

Characteristic	Category	Sample n	Percent
Gender	Male	7	5.83%
	Female	112	93.33%
	Diverse	1	0.83%
Degree	Bachelor	116	96.67%
	Master	3	2.50%
	Postgraduate	1	0.83%

Figure 3 summarizes which AI technologies are used by the study participants. The most frequently used type of AI technology was text-generating AI chatbots, followed by voice assistants like Siri or Alexa.

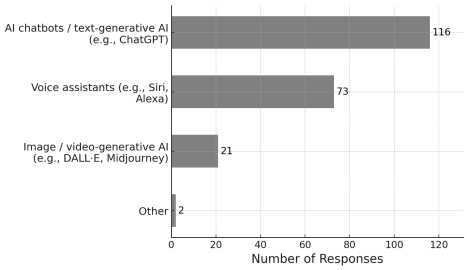


Figure 3. Types of AI technologies used (multiple answers allowed).

Figure 4 shows that study participants employed LLMs most frequently in academic contexts.

E. Instruments

We used validated questionnaires to measure the key constructs underlying our research questions. All Likert scales used 1-5 response options unless otherwise noted.

1) Personality: To assess Extraversion and Neuroticism, we used the Big Five Inventory (BFI-44) [32], each trait captured with 8 items on a 5-point Likert scale. As we did not use the full questionnaire, we conducted a reliability analysis. The original Extraversion subscale showed insufficient reliability

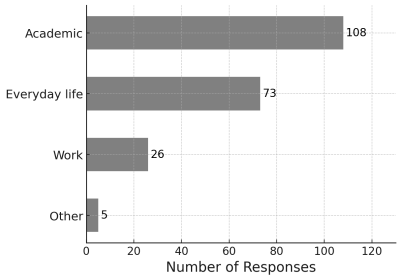


Figure 4. Context of LLM Usage (multiple answers allowed).

(Cronbach's $\alpha = .353$). Two items with negative item–total correlations were removed, resulting in a 6-item Extraversion scale. The Neuroticism subscale already demonstrated satisfactory internal consistency, so no items were excluded. For the final sample, internal consistencies were $\alpha = .849$ for Extraversion and $\alpha = .860$ for Neuroticism, indicating good reliability.

2) LLM Usage: To assess LLM Usage, we developed a new subscale based on MTUAS [27]. We reduced the original ten-point frequency scale to nine options due to redundancy in the highest categories. Items covered typical application contexts, such as information search, creative writing, academic text editing, study support, and rapid problem-solving [33][34]. The subscale showed high internal consistency ($\alpha = .890$). An exploratory factor analysis confirmed a one-factor structure ($KMO = .844$; Bartlett's test $\chi^2(10) = 363.49$, $p < .001$).

3) Technology Acceptance: Technology Acceptance was measured with a 13-item scale adapted from the DTAS [28], a validated German instrument based on the classical TAM framework. Items were randomized to avoid response patterns and answered on a five-point Likert scale. The scale demonstrated excellent internal consistency in the present sample ($\alpha = .925$).

4) Cognitive Dissonance: To the best of our knowledge, no established instrument exists to measure Cognitive Dissonance in the context of LLMs. We therefore adapted the well-established scale by Sweeney et al. [35], originally developed for purchase decisions, which can be contextually adapted to LLM Usage without loss of content structure. The 22 items cover three dimensions: emotions, wisdom of purchase, and concern over deal. Therefore, we calculated an overall average for all items. To preserve this thematic grouping, items were presented in thematic blocks rather than fully randomized. We provided an instructional text with a concrete example to enhance response quality. All items were rated on a five-point Likert scale. The adapted scale showed high internal consistency in our sample (Cronbach's $\alpha = .894$).

F. Assumption Testing

To test **H1** and **H2**, we conducted correlation analyses between Extraversion or Neuroticism and Technology Acceptance or Cognitive Dissonance. Normality, assessed with Shapiro–Wilk tests, was not met. Given the ordinal scaling of the variables and the sample size, Spearman's correlation was used.

For **H3a**, we tested a mediation model with Neuroticism (independent variable), Cognitive Dissonance (mediator), and Technology Acceptance (dependent variable) following [36]. Because the mediator and dependent variables deviated from normality in Shapiro–Wilk tests, we used bootstrap mediation with 5,000 samples and 95% CIs, which does not assume normality and provides robust estimates. Effects were considered significant if the interval did not include zero. As Extraversion was unrelated to technology acceptance and LLM usage, **H3b** was treated as exploratory.

The exploratory hypotheses **H4** and **H5** were tested analogously to **H1** and **H2** using Spearman's rank correlations, as assumptions for parametric tests were not met.

IV. Results

To evaluate our hypotheses, we first present the descriptive statistics of the central variables in Table III. Extraversion and Neuroticism are centered around the scale midpoints with moderate variance. Technology Acceptance ratings were generally positive, while Cognitive Dissonance was reported at moderate levels. LLM Usage varied considerably, with most participants reporting occasional use, but a small subgroup indicating high-frequency engagement.

TABLE III. Descriptive statistics of main variables.

Variable	Mean (SD)	Range
Extraversion	3.21 (0.74)	1–5
Neuroticism	2.99 (0.73)	1–5
Technology Acceptance	3.63 (0.73)	1–5
Cognitive Dissonance	2.11 (0.67)	1–5
LLM Usage	4.05 (1.94)	1–9

A. Extraversion and Technology Acceptance (*H1*)

Contrary to our expectations, Extraversion did not predict higher acceptance of LLMs. The correlation between Extraversion and Technology Acceptance was very weak and non-significant ($r_s = .061$, $p = .509$). Thus, **H1** could not be confirmed. As shown in Figure 5, no systematic patterns emerge.

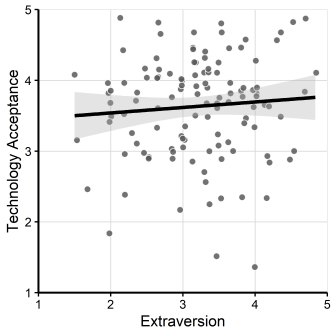


Figure 5. Correlation of Extraversion and Technology Acceptance.

These findings suggest that sociability and outgoingness do not substantially influence whether students perceive LLMs as useful or easy to use.

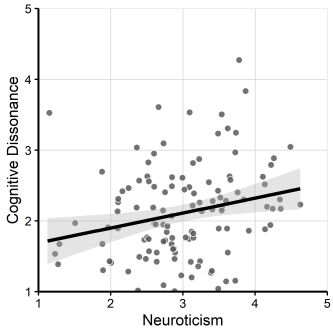


Figure 6. Correlation of Cognitive Dissonance and Neuroticism.

B. Neuroticism and Cognitive Dissonance (*H2*)

The analysis revealed a small-to-moderate, statistically significant positive association between Neuroticism and Cognitive Dissonance in the expected direction ($r_s = .243$, $p = .008$). Thus, the results provide statistical evidence for a relationship between Neuroticism and the experience of Cognitive Dissonance when using LLMs in the present sample.

Figure 6 illustrates the positive association between Neuroticism and Cognitive Dissonance.

C. Mediation of Cognitive Dissonance between Neuroticism and Technology Acceptance (*H3a*)

Neuroticism significantly predicted Cognitive Dissonance (estimate $= 0.315$, $SE = 0.121$, $p = .009$, 95%-CI [0.057; 0.558]) and Cognitive Dissonance positively predicted Technology Acceptance (estimate $= 0.225$, $SE = 0.090$, $p = .013$, 95%-CI [0.023; 0.448]). Bootstrapping indicated a small indirect effect (estimate $= 0.071$, $SE = 0.039$, 95%-CI [0.005; 0.212]) that was significant by our CI criterion (95%-CI excluded 0), although the associated p-value was slightly above .05 ($p = .072$). The direct effect of Neuroticism on Technology Acceptance remained negative and significant (estimate $= -0.293$, $SE = 0.121$, $p = .017$, 95%-CI [–0.505; –0.047]), indicating a partial mediation. These results suggest that Cognitive Dissonance partly mediates the relationship between Neuroticism and Technology Acceptance of LLMs.

Figure 7 provides an overview of the mediation model linking Neuroticism, Cognitive Dissonance, and Technology Acceptance of LLMs.

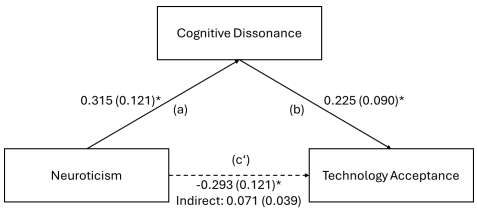


Figure 7. Coefficients and standard errors from bootstrapped mediation analysis of Neuroticism, Cognitive Dissonance, and Technology Acceptance. * indicates $p < .05$, ** indicates $p < .001$.

*D. Mediation of Technology Acceptance between Extraversion and LLM Usage (**H3b**)*

Extraversion neither predicted LLM Usage (estimate = 0.016, $SE = 0.095$, $p = .869$, 95%-CI [–0.158; 0.174]) nor Technology Acceptance (estimate = 0.106, $SE = 0.123$, $p = .388$, 95%-CI [–0.149; 0.353]), whereas Technology Acceptance strongly predicted LLM Usage (estimate = 0.642, $SE = 0.070$, $p < .001$, 95%-CI [0.530; 0.758]). The indirect effect of Extraversion via Technology Acceptance was non-significant (estimate = 0.068, $SE = 0.079$, $p = .390$, 95%-CI [–0.100; 0.226]).

Figure 8 illustrates the mediation model that links Extraversion, Technology Acceptance, and LLM Usage.

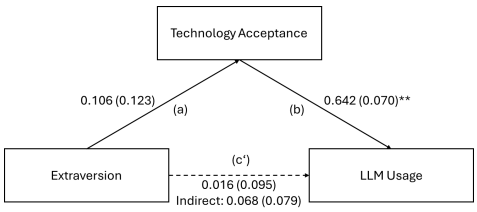


Figure 8. Coefficients and standard errors from bootstrapped mediation analysis of Extraversion, Technology Acceptance, and LLM Usage. * indicates $p < .05$, ** indicates $p < .001$.

*E. Neuroticism and Technology Acceptance (**H4**)*

Spearman's correlation indicated a weak, non-significant negative association between Neuroticism and Technology Acceptance ($r_s = -.157$). The significance level ($p = .087$) was slightly above the threshold of 0.05. This suggests only a small trend that might reach significance in larger samples, with Neuroticism explaining little variance in Acceptance. Figure 9 visualizes this weak negative trend.

The regression line indicates a small negative relationship.

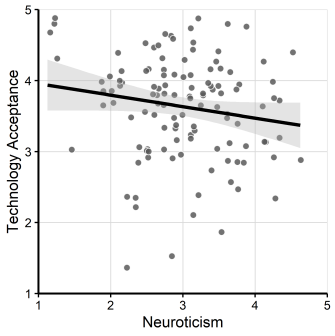


Figure 9. Correlation of Neuroticism and Technology Acceptance.

*F. Extraversion and Cognitive Dissonance (**H5**)*

Extraversion showed a weak, non-significant negative correlation with Cognitive Dissonance ($r_s = -.136$, $p = .138$), indicating no meaningful association, which is also reflected in Figure 10. The effect size ($z = -.137$) was smaller than the associated standard error ($SE = 0.093$), which further indicated low stability of the effect.

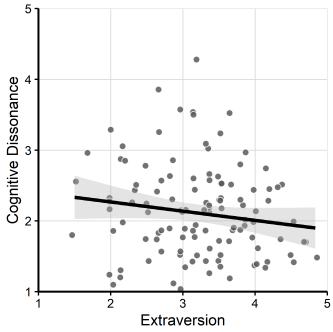


Figure 10. Correlation of Extraversion and Cognitive Dissonance.

V. Discussion

Our results provide first indications that are relevant for both theory and design. The unexpected mediation might suggest a mechanism-level explanation in which Cognitive Dissonance helps understand why neuroticism could be related to higher LLM acceptance. In contrast, the non-significant result for **H1** could indicate a possible boundary condition for extraversion in this context.

A. Interpretation of Main Findings

H1 was not supported: Extraversion was not related to Technology Acceptance of LLMs. This contrasts with prior work that found Extraversion to be a key predictor of Technology Acceptance [23][37] and may partly reflect our specific sample of psychology students.

H2 revealed a positive correlation between Neuroticism and the experience of Cognitive Dissonance when using LLMs. This confirms our hypothesis, even though the effect size was small. The findings suggest that psychology students with higher levels of Neuroticism tend to experience stronger Cognitive Dissonance when interacting with LLMs. They report more intense negative emotions when disappointed by the technology [25]. Because Neuroticism leads to heightened uncertainty, anxiety, and emotionality, individuals high in this trait are more likely to react strongly when faced with contradictions or uncertainty [24]. Our finding is consistent with prior studies indicating that Neuroticism may predict negative emotions, uncertainty, and even avoidance of LLM use [38].

Cognitive Dissonance was generally low to moderate, suggesting that dissonance in LLM use is not universal. Other individual factors (e.g., self-efficacy, cognitive load, and learning motivation) likely contribute to how students perceive LLMs [39].

Surprisingly, H3a revealed a positive and significant mediation effect based on bootstrap confidence intervals, suggesting that higher levels of Neuroticism are associated with higher Technology Acceptance through the experience of Cognitive Dissonance, rather than lower acceptance, as hypothesized. Festinger's theory of dissonance reduction offers a possible explanation: Experiencing Cognitive Dissonance creates psychological tension, which individuals reduce by adjusting their

attitudes and behaviors [15]. When expectations regarding LLMs are not met, users may feel anger, frustration, or tension [29]. These negative feelings can be reduced by a change in attitude. This could, for example, be achieved by accepting the technology as a helpful tool in their studies. Coping strategies for dissonance may thus promote acceptance [25]. The positive mediation effect is also in line with the theory of transformative education, which states that learning can be triggered by the experience of irritation and otherness [40]. Both direct and indirect effects between Neuroticism and Technology Acceptance were significant, indicating partial mediation, with Cognitive Dissonance explaining only part of the relationship.

Extraversion showed no significant associations with technology acceptance or LLM usage; therefore, **H3b** was not supported. The direct effect of Extraversion on Technology Acceptance (as suggested in **H1**) was also not significant. The path between Technology Acceptance and LLM Usage was highly significant. Higher acceptance was linked to higher frequency of use among our subjects. However, no direct significant relationship was found between Extraversion and either acceptance or LLM Usage. Thus, Extraversion appears unrelated to the investigated variables. We conclude that while Extraversion does not influence Technology Acceptance of psychology students, Technology Acceptance itself positively predicts LLM Usage, driven by perceptions of usefulness and ease of use [8].

Relatedly, adoption and continued use may depend on users' perceived capability to handle the tool effectively. Differences in digital/AI self-efficacy, AI literacy, and task-related cognitive demands could affect both the experience of dissonance and acceptance and may shape the observed relationships.

B. Interpretation of Exploratory Findings

Exploratory analyses for **H4** suggested a small negative association between Neuroticism and Technology Acceptance. However, the effect was weak and explained only $\approx 2.5\%$ of the variance. While some prior work reports similar negative trends [10], other studies in the context of ChatGPT find no clear relationship [38]. Overall, Neuroticism alone appears to play only a minor role in explaining differences in acceptance [38].

For **H5**, we found no meaningful relationship between Extraversion and Cognitive Dissonance in LLM use; the effect was very small ($\approx 2\%$ explained variance) and statistically negligible.

C. Implications for HCI and Educational Practice

Based on our findings, several implications arise for the design of LLM-supported learning environments in higher education. Students higher in Neuroticism appear more sensitive to negative outcomes when using LLMs. Universities can support responsible use by calibrating expectations about capabilities and limitations, providing transparent rationales and uncertainty cues, scaffolding error recovery with reversible actions and safe defaults, and embedding brief reflective prompts after mismatches. Such measures may reduce emotional strain while promoting informed Technology Acceptance.

The unexpected mediation result, in which Cognitive Dissonance can, under specific conditions, contribute to higher Technology Acceptance, suggests that dissonant experiences can sometimes be turned into learning opportunities when adequately supported. Personality-sensitive design may therefore help tailor feedback and guidance to learners who are particularly vulnerable to contradictory or erroneous information. Personalized scaffolds that reduce Cognitive Dissonance and targeted support for students with higher levels of Neuroticism could help stabilize constructive engagement with these tools.

D. Limitations

Several limitations need to be considered when interpreting our findings.

First, the study relied on a homogeneous sample of psychology students from a single university, most of whom were enrolled in distance study programs. Moreover, most of the participants were females. This focus naturally limits the generalizability of the results to broader populations.

Second, the study followed a correlational, cross-sectional design, which precludes causal conclusions.

Third, Cognitive Dissonance, Technology Acceptance, and LLM Usage were measured with adapted self-report scales that had not originally been validated in the specific context of LLMs. This may reduce construct validity and makes the results susceptible to influences like social desirability or response biases.

Finally, personality was assessed using a shortened version of the BFI-44 that included only the traits Neuroticism and Extraversion, and two Extraversion items were removed to improve reliability. While this yielded acceptable internal consistencies, it narrows content coverage and omits other potentially relevant traits from the Big Five model.

VI. Conclusion and Future Work

In summary, our study contributes to understanding the role of personality in shaping Technology Acceptance of LLMs in higher education. Neuroticism showed small negative associations with acceptance, including a significant negative direct effect in the mediation model. Cognitive Dissonance played a small but theoretically interesting mediating role, indicating that affective reactions may shape acceptance in nuanced ways

Our results highlight the importance of individual differences in human-LLM interaction. Although this interface remains underexplored, our findings underline its relevance from a user perspective and provide a basis for designing adaptive, user-centered AI technologies in educational contexts. Within our student sample, LLM Usage appears to be driven primarily by Technology Acceptance rather than by the examined personality traits, and Neuroticism functions as an indicator for experiencing Cognitive Dissonance rather than as a direct driver of usage.

Our findings yield several implications for future research.

First, longitudinal and experimental studies (e.g., manipulating feedback quality or error frequency) are needed to clarify causal relationships between Neuroticism, Cognitive Dissonance, and Technology Acceptance.

Second, future work should examine more diverse and representative populations beyond psychology students to assess the generalizability of the observed patterns across disciplines and educational settings. In addition, future studies could examine whether group-level adoption climates (e.g., cohort norms or institutional AI policies) aligned with TAM2's social influence mechanisms account for variance beyond individual traits by using multilevel designs. Comparative research across technology classes (e.g., smartphones vs. LLM-based tools) could help distinguish general acceptance mechanisms (e.g., perceived usefulness and ease of use) from effects that are specific to generative AI (e.g., uncertainty and accountability demands due to lack of transparency).

Third, additional Big Five traits (e.g., Openness, Conscientiousness, Agreeableness) and domain-specific, validated instruments should be incorporated to obtain more nuanced assessments of Cognitive Dissonance and LLM Usage. Future work should also include capability-related measures such as AI literacy and digital self-efficacy, and test whether cognitive demands (e.g., task complexity) moderate links between personality, dissonance, and acceptance.

Finally, qualitative and mixed-methods approaches could provide richer insights into users' subjective experiences and emotional reactions, helping to identify specific triggers of dissonance that may not surface in quantitative measures.

REFERENCES

[1] S. A. M. Aldosari, "The Future of Higher Education in the Light of Artificial Intelligence Transformations", *International Journal of Higher Education*, vol. 9, no. 3, p. 145, Mar. 2020, ISSN: 1927-6052, 1927-6044. DOI: 10.5430/ijhe.v9n3p145.

[2] A. Casheekar, A. Lahiri, K. Rath, K. S. Prabhakar, and K. Srinivasan, "A contemporary review on chatbots, AI-powered virtual conversational agents, ChatGPT: Applications, open challenges and future research directions", *Computer Science Review*, vol. 52, p. 100 632, May 2024, ISSN: 15740137. DOI: 10.1016/j.cosrev.2024.100632.

[3] A. Watanabe, T. Schmohl, and K. Schelling, "Akzeptanzforschung zum Einsatz Künstlicher Intelligenz in der Hochschulbildung. Eine kritische Bestandsaufnahme [acceptance research on the use of artificial intelligence in higher education: A critical review]", in *Künstliche Intelligenz in Der Bildung*, C. de Witt, C. Gloerfeld, and S. E. Wrede, Eds., Wiesbaden: Springer Fachmedien Wiesbaden, 2023, pp. 263–289, ISBN: 978-3-658-40079-8. DOI: 10.1007/978-3-658-40079-8_13.

[4] A. Baillifard, M. Gabella, P. B. Lavenex, and C. S. Martarelli, *Implementing Learning Principles with a Personal AI Tutor: A Case Study*, 2023. DOI: 10.48550/arXiv.2309.13060.

[5] R. Mehlan, C. Hess, Q. Stierstorfer, and K. Schaaff, "Personalized Knowledge Transfer Through Generative AI: Contextualizing Learning to Individual Career Goals", in *Artificial Intelligence in Education Technologies: New Development and Innovative Practices*, T. Schlippe, E. C. K. Cheng, and T. Wang, Eds., Singapore: Springer Nature Singapore, 2026.

[6] E. Kasneci et al., "ChatGPT for good? On opportunities and challenges of large language models for education", *Learning and Individual Differences*, vol. 103, p. 102 274, Apr. 2023, ISSN: 10416080. DOI: 10.1016/j.lindif.2023.102274.

[7] M. Bernabei, S. Colabianchi, A. Falegnami, and F. Costantino, "Students' Use of Large Language Models in Engineering Education: A Case Study on Technology Acceptance, Perceptions, Efficacy, and Detection Chances", *Computers and Education: Artificial Intelligence*, vol. 5, p. 100 172, 2023, ISSN: 2666920X. DOI: 10.1016/j.caeai.2023.100172.

[8] F. D. Davis, "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology", *MIS Quarterly*, vol. 13, no. 3, pp. 319–340, Sep. 1989, ISSN: 02767783. DOI: 10.2307/249008.

[9] V. Venkatesh and F. Davis, "A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies", *Management Science*, vol. 46, pp. 186–204, Feb. 2000. DOI: 10.1287/mnsc.46.2.186.11926.

[10] R. Riedl, "Is Trust in Artificial Intelligence Systems Related to User Personality? Review of Empirical Evidence and Future Research Directions", *Electronic Markets*, vol. 32, no. 4, pp. 2021–2051, Dec. 2022, ISSN: 1019-6781, 1422-8890. DOI: 10.1007/s12525-022-00594-4.

[11] C. Calluso and M. G. Devetag, "The Impact of Technology Acceptance and Personality Traits on the Willingness to Use AI-assisted Hiring Practices", *International Journal of Organizational Analysis*, vol. 33, no. 5, pp. 1368–1385, Jun. 2025, ISSN: 1934-8835, 1758-8561. DOI: 10.1108/ijoa-06-2024-4562.

[12] F. D. O. Santini et al., "Understanding Students' Technology Acceptance Behaviour: A Meta-Analytic Study", *Technology in Society*, vol. 81, p. 102 798, Jun. 2025, ISSN: 0160-791X. DOI: 10.1016/j.techsoc.2024.102798.

[13] H. Holdefehr, M.-A. Heidelmann, and K. Schaaff, "Ai use in the workplace: Correlational evidence on motivation, autonomy, job security, and ai-related threat", in *COGNITIVE 2026, The Eighteenth International Conference on Advanced Cognitive Technologies and Applications*, 2026.

[14] Z. Wen et al., *Self-assessment, Exhibition, and Recognition: A Review of Personality in Large Language Models*, Jun. 2024. DOI: 10.48550/arXiv.2406.17624.

[15] L. Festinger, *A Theory of Cognitive Dissonance*. Stanford, CA: Stanford University Press, 1957, ISBN: 0804709114.

[16] S. A. D. Popenici and S. Kerr, "Exploring the Impact of Artificial Intelligence on Teaching and Learning in Higher Education", *Research and Practice in Technology Enhanced Learning*, vol. 12, no. 1, p. 22, Dec. 2017, ISSN: 1793-7078. DOI: 10.1186/s41039-017-0062-8.

[17] A. S. Almogren, W. M. Al-Rahmi, and N. A. Dahri, "Exploring Factors Influencing the Acceptance of ChatGPT in Higher Education: A Smart Education Perspective", *Heliyon*, vol. 10, no. 11, e31887, 2024, ISSN: 2405-8440. DOI: 10.1016/j.heliyon.2024.e31887.

[18] T. Kühbacher, T. Schlippe, and K. Schaaff, "Which Chatbot Is the Most Empathic Teacher?", in *Artificial Intelligence in Education Technologies: New Development and Innovative Practices*, T. Schlippe, E. C. K. Cheng, and T. Wang, Eds., Singapore: Springer Nature Singapore, 2025, pp. 56–73, ISBN: 978-981-97-9255-9.

[19] R. R. McCrae and P. T. Costa Jr., "A Five-Factor Theory of Personality.", in *Handbook of Personality: Theory and Research, 2nd Ed.* New York, NY, US: Guilford Press, 1999, pp. 139–153, ISBN: 1-57230-483-9 (Hardcover).

[20] A. Ghimire and J. Edwards, *Generative AI Adoption in Classroom in Context of Technology Acceptance Model (TAM) and the Innovation Diffusion Theory (IDT)*, Mar. 2024. DOI: 10.48550/arXiv.2406.15360.

[21] R. Kanai and G. Rees, "The Structural Basis of Inter-Individual Differences in Human Behaviour and Cognition", *Nature Reviews Neuroscience*, vol. 12, no. 4, pp. 231–242, Apr. 2011, ISSN: 1471-003X, 1471-0048. DOI: 10.1038/nrn3000.

[22] R. R. McCrae and P. T. Costa, "Validation of the Five-Factor Model of Personality across Instruments and Observers.", *Journal of Personality and Social Psychology*, vol. 52, no. 1, pp. 81–90, 1987, ISSN: 1939-1315, 0022-3514. DOI: 10.1037/0022-3514.52.1.81.

[23] D. Seibert, A. Godulla, and C. Wolf, *Understanding How Personality Affects the Acceptance of Technology: A Literature Review*. Leipzig, 2021, p. 24.

[24] J. B. Hirsh and M. Inzlicht, "The Devil You Know: Neuroticism Predicts Neural Response to Uncertainty", *Psychological Science*, vol. 19, no. 10, pp. 962–967, Oct. 2008, ISSN: 0956-7976, 1467-9280. DOI: 10.1111/j.1467-9280.2008.02183.x.

[25] D. Marikyan, S. Papagiannidis, and E. Alamanos, "When Technology Does Not Meet Expectations: A Cognitive Dissonance Perspective", in *UK Academy for Information Systems 2020*, United States: Association for Information Systems, Aug. 2020.

[26] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User Acceptance of Information Technology: Toward a Unified View", *MIS Quarterly*, vol. 27, no. 3, pp. 425–478, 2003, ISSN: 02767783. DOI: 10.2307/30036540.

[27] L. Rosen, K. Whaling, L. Carrier, N. Cheever, and J. Rokkum, "The Media and Technology Usage and Attitudes Scale: An empirical investigation", *Computers in Human Behavior*, vol. 29, no. 6, pp. 2501–2511, Nov. 2013, ISSN: 07475632. DOI: 10.1016/j.chb.2013.06.006.

[28] A. Schorr, "Skala zur Erfassung der Digitalen Technologieakzeptanz – Weiterentwicklung zum testtheoretisch geprüften Instrument [scale for measuring digital technology acceptance. further development and testing of the scale based on classical testing theory]", in *Digitale Arbeit, digitaler Wandel, digitaler Mensch? 66. Kongress für Arbeitswissenschaft*, G. für Arbeitswissenschaft, Ed., Dortmund: GfA-Press, 2020, pp. 1–7.

[29] C. M. Montes and R. Khojah, *Emotional Strain and Frustration in LLM Interactions in Software Engineering*, Apr. 2025. DOI: 10.48550/arXiv.2504.10050.

[30] M. G. Dawson, R. Deer, and S. Boguslawski, "Cognitive Dissonance in Programming Education: A Qualitative Exploration of the Impact of Generative AI on Application-Directed Learning", *Computers in Human Behavior Reports*, vol. 19, p. 10, 2025, ISSN: 2451-9588. DOI: 10.1016/j.chbr.2025.100724.

[31] M. Mondal, L. Dolamic, G. Bovet, and P. Cudre-Mauroux, *Do Large Language Models Exhibit Cognitive Dissonance? Studying the Difference Between Revealed Beliefs and Stated Answers*, Jun. 2024. DOI: 10.48550/arXiv.2406.14986.

[32] O. P. John and S. Srivastava, "The Big Five trait taxonomy: History, measurement, and theoretical perspectives", in *Handbook of Personality: Theory and Research*, L. A. Pervin and O. P. John, Eds., vol. 2, New York, NY, USA: Guilford Press, 1999, pp. 102–138.

[33] K. Guo and D. Li, "Understanding EFL Students' Use of Self-Made AI Chatbots as Personalized Writing Assistance Tools: A Mixed Methods Study", *System*, vol. 124, p. 103 362, Aug. 2024, ISSN: 0346-251X. DOI: 10.1016/j.system.2024.103362.

[34] L. Labadze, M. Grigolia, and L. Machaidze, "Role of AI chatbots in education: Systematic literature review", *International Journal of Educational Technology in Higher Education*, vol. 20, no. 1, Oct. 2023, ISSN: 2365-9440. DOI: 10.1186/s41239-023-00426-1.

[35] J. C. Sweeney, D. R. Hausknecht, and G. N. Soutar, "Cognitive Dissonance after Purchase: A Multidimensional Scale.", *Psychology & Marketing*, vol. 17, pp. 369–385, 2000.

[36] R. M. Baron and D. A. Kenny, "The Moderator–Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations.", *Journal of Personality and Social Psychology*, vol. 51, no. 6, pp. 1173–1182, 1986, ISSN: 1939-1315, 0022-3514. DOI: 10.1037/0022-3514.51.6.1173.

[37] E. S. D. Duro, G. A. Veltri, H. Golino, and M. Stella, *Measuring and Identifying Factors of Individuals' Trust in Large Language Models*, Mar. 2025. DOI: 10.48550/arXiv.2502.21028.

[38] J. De Winter, D. Dodou, and Y. B. Eisma, "Personality and Acceptance as Predictors of ChatGPT Use", *Discover Psychology*, vol. 4, no. 1, May 2024, ISSN: 2731-4537. DOI: 10.1007/s44202-024-00161-2.

[39] S. Lambiase, G. Catolino, F. Palomba, F. Ferrucci, and D. Russo, *Exploring Individual Factors in the Adoption of LLMs for Specific Software Engineering Tasks*, Apr. 2025. DOI: 10.48550/arXiv.2504.02553.

[40] H.-C. Koller, *Bildung anders denken: Einführung in die Theorie transformatorischer Bildungsprozesse [Thinking education differently: Introduction to the theory of transformational educational processes]*, 3., erweiterte und aktualisierte Auflage. Stuttgart: Verlag W. Kohlhammer, 2023, ISBN: 978-3-17-042795-2.

AI Use in the Workplace: Correlational Evidence on Motivation, Autonomy, Job Security, and AI-Related Threat

Hannah Holdefehr, Marc-André Heidelmann, Kristina Schaaff
IU International University of Applied Sciences, Erfurt, Germany
e-mail: {marc-andre.heidelmann | kristina.schaaff}@iu.org

***Abstract*—In our study, we investigate how artificial intelligence (AI) use relates to key psychological factors at work: motivation and job satisfaction, perceived autonomy, and job security. In a cross-sectional survey, 185 employees from multiple industries completed a combination of validated, adapted, and self-developed measures on AI use frequency, perceived autonomy, perceived AI-related threat, intrinsic and extrinsic motivation, and Big Five personality traits. AI use was positively associated with autonomy and intrinsic motivation, while effects on extrinsic motivation were smaller. However, as the autonomy indicator and the extrinsic motivation subscale showed low internal consistency in this sample, findings involving these constructs should be interpreted cautiously. Openness to experience predicted higher AI use. Neuroticism and extraversion were linked to more serious perceived AI-related threats, whereas conscientiousness positively predicted perceived autonomy in AI-supported work. Overall, the findings highlight the role of personality in shaping workplace AI use and AI-related perceptions and its psychological and job-security–related consequences.**

Keywords-AI in the workplace; employee motivation; job satisfaction; AI use and personality

I. INTRODUCTION

The growing integration of AI into organizational processes is reshaping workflows, redistributing tasks, and transforming employees' psychological experiences at work [1]. While organizations increasingly rely on AI to enhance efficiency and productivity, employees frequently express uncertainty regarding loss of control, job security, and shifts in role expectations. Recent industry cases suggest AI is already restructuring administrative work (e.g., [2][3]).

Despite this rapid transformation, the psychological consequences of AI use—particularly its effects on motivation, autonomy, and perceived threat—remain insufficiently explored [4][5]. Existing research has primarily focused on economic and technical indicators, whereas employees' subjective experiences and motivational processes have received comparatively little empirical attention.

In our study, we address this research gap by examining how AI use is associated with central psychological constructs in the workplace and how individual personality traits shape these perceptions. Our objective is to analyze the associations between workplace AI use and employees' motivation, autonomy, and job security. The investigation builds on Self-Determination Theory (SDT) by Ryan and Deci [5], as well as technology-acceptance frameworks [6]. Additionally, the predicting role of personality traits, operationalized via the Big Five model [7], is considered.

The study addresses four central research questions:

- RQ1: How is AI use associated with employees' perceptions of job security and autonomy?
- RQ2: To what extent is AI use associated with intrinsic and extrinsic motivation as well as overall job satisfaction?
- RQ3: Do the effects of workplace AI use on employees' intrinsic and extrinsic motivation differ across industries or occupational groups?
- RQ4: How do individual personality traits (Big Five) shape employees' perceived AI-related threat, perceived autonomy, and actual AI use in the workplace?

All four research questions were specified prior to data analysis and treated as exploratory-correlational, without formal confirmatory hypotheses.

Existing studies on AI in the workplace have mostly focused on productivity, technology acceptance, or single psychological variables in isolation. Autonomy, intrinsic and extrinsic motivation, job security, and Big Five personality traits have rarely been addressed within one coherent model. Our contributions are as follows:

1) We jointly analyze how AI use relates to these psychological outcomes,
2) We integrate personality traits as predictors of AI use and perceived AI-related threat, and
3) We derive practical implications for AI-based work design that explicitly consider individual differences.

The remainder of this paper is structured as follows. In Section II, we review the theoretical background In Section III, we describe the study overview and design, including the research approach, sample, instruments, quality criteria, and statistical analyses. In Section IV, we present the empirical results, covering descriptive findings, exploratory analyses of AI use, associations with job security, autonomy, motivation, and job satisfaction, as well as differences across occupational groups and personality-based predictors. In Section V, we discuss the findings in relation to existing research, derive implications for organizations and work design, and outline the study's limitations. Finally, Section VI concludes the paper and highlights directions for future research.

II. BACKGROUND

This section outlines the theoretical foundations of the study by reviewing prior research on AI in the workplace, employee motivation and job satisfaction, job security and autonomy, and the role of personality traits in shaping employees' responses to technological change.

A. Artificial Intelligence in the Workplace

In organizational contexts, AI increasingly takes over tasks previously performed by human experts, reshaping work processes, skill requirements, and professional roles [8]. Current research points to ambivalent effects: AI can accelerate workflows and reduce errors, yet employees often report uncertainty, especially when systems are used for performance monitoring or automated decision-making [9]. Lack of transparency or insufficient employee involvement can reduce trust and negatively affect job satisfaction and technology acceptance [10]. Recent studies also show that employees sometimes perceive AI as a social interaction partner, which can foster emotional attachments but also psychological strain, such as feelings of surveillance, comparison with algorithmic standards, or unclear boundaries between human and machine agency [11]. Overall, these findings suggest that the effects of AI depend less on the technology itself and more on employees' subjective interpretations and the organizational conditions surrounding implementation.

B. Employee Motivation and Job Satisfaction

Work psychology offers several frameworks for understanding how AI influences employees' motivation. SDT [5], the Unified Theory of Acceptance and Use of Technology (UTAUT) [6], and Herzberg's Two-Factor Theory [12] all emphasize that motivational outcomes depend on whether AI supports or undermines autonomy, competence, meaningfulness, and perceived support. AI can enhance motivation when it increases control, task variety, and competence, but may reduce motivation and satisfaction when used primarily for monitoring or opaque algorithmic decision-making [13]–[15]. Although the literature acknowledges these benefits and risks, systematic empirical studies on the underlying psychological mechanisms remain scarce, with prior work focusing largely on economic or efficiency outcomes [16].

C. Job Security and Autonomy

Job security and autonomy are central psychological constructs shaping responses to technological change. In the context of AI, perceived job security depends less on objective job-loss risks than on subjective interpretations of control, competence, and expected change [16]; AI applications that automate decisions or monitor performance are particularly associated with declines in perceived job security [9]. Autonomy is a key predictor of motivation and job satisfaction and is highly sensitive to technological interventions. Algorithmic systems can reduce perceived autonomy even when efficiency improves, but may enhance it when they remove routine tasks and free cognitive and temporal resources [13]. A decisive factor is whether AI is perceived as supportive or controlling, which directly shapes motivation, satisfaction, and acceptance [5]. Organizational context further influences experiences of uncertainty and control: employees in smaller or less digitally mature organizations often report higher overload and insufficient communication, intensifying perceived threat [15]. Thus, job security and autonomy are jointly shaped by individual perceptions and organizational implementation and communication practices.

D. Personality Traits and the Big Five Model

Personality psychology assumes that stable interindividual differences shape how people evaluate and respond to their environment. The Big Five model, one of the most established frameworks, describes personality along openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism [7]. In the work context, these traits correlate with attitudes, satisfaction, and behavior; Judge et al. identified the Big Five—particularly neuroticism and conscientiousness—as strong predictors of job satisfaction [17]. Technology-acceptance models similarly suggest that individual dispositions influence whether new technologies are perceived as helpful or threatening [6]. Research on digital technologies supports these associations: openness is linked to greater willingness to adopt new technologies, whereas individuals high in neuroticism react more strongly to uncertainty and evaluate technological change more critically [18]. However, there is also work showing no association between extraversion and technology acceptance, while neuroticism had a negative direct effect [19]. Despite this relevance, personality differences in the context of workplace AI have received limited empirical attention, making the Big Five a valuable framework for explaining why employees differ in their perceptions, attitudes, and reactions toward AI-based systems.

III. Study Overview and Design

This section presents the methodological foundation of the study, including the overall research approach, sample and data collection, operationalization of key constructs, quality criteria of the measures, and the statistical analyses used to address the research questions.

A. Research Approach and Objectives

In our study, we examine associations between workplace AI use and employees' motivation, perceived autonomy, perceived job security, and perceived AI-related threat, and the predictive role of Big Five personality traits [7]. Using a quantitative cross-sectional survey grounded in SDT and UTAUT [6], we capture employee perceptions at a single point in time.

B. Sample and Data Collection

Data were collected via a standardized online survey distributed through social networks, professional networks, and email distribution lists. The final sample included 185 employed adults ($\geq$ 18 years) from diverse industries and occupational groups. Data quality was ensured by screening for completeness, identifying outliers, and excluding inattentive respondents. No occupational groups were excluded, ensuring broad representation of professional backgrounds, levels of experience, and work contexts. Validated and widely used psychological scales were adapted for this study. Key variables included frequency of AI use, perceived autonomy, perceived AI-related threat, intrinsic and extrinsic motivation, and personality traits

according to the Big Five. Data collection via an online platform ensured a standardized administration procedure and efficient participation. The survey structure was designed to capture self-reported indicators of AI use and subjective psychological responses.

C. Instruments and Operationalization

A combination of validated psychological scales and self-developed items was used to assess the key constructs, with adaptations made where necessary to fit the context of AI-supported work. AI use was assessed with a self-developed scale capturing both frequency and type of AI-supported activities (e.g., administrative tasks, decision support). Motivation was measured through established scales of intrinsic and extrinsic motivation as well as job satisfaction [5]. Job satisfaction was assessed using a validated single-item measure on a 5-point Likert scale (1 = very dissatisfied, 5 = very satisfied) and analyzed as an outcome distinct from motivation . Autonomy and perceived job security were assessed using self-developed items created for the original study.

Personality traits were measured using an established Big Five inventory, capturing the five broad personality dimensions: extraversion, neuroticism, openness to experience, conscientiousness, and agreeableness [7][17]. Additional demographic variables, including age, years of professional experience, and sector, were collected to support contextual interpretation and further exploratory analyses.

D. Quality Criteria

To ensure the methodological rigor of the study, the measurement instruments were evaluated according to classical test quality criteria: objectivity, reliability, and validity.

Objectivity was considered high due to the standardized online administration via Microsoft Forms. All participants received identical instructions and responded to the same items in the same order. Automated answer coding, predefined recording rules, and the use of statistical software ensured scoring and evaluation objectivity, consistent with the procedures outlined in the original study [20].

Reliability was assessed using Cronbach's alpha as an indicator of internal consistency. The job security scale demonstrated acceptable reliability ($\alpha = .73$, $\omega = .76$). However, the autonomy scale showed weaker internal consistency ($\alpha = .32$, $\omega = .49$), suggesting potential limitations in item quality or construct coverage [21]. This limitation should be addressed in future studies, as it may reduce precision in measuring autonomy. Intrinsic motivation showed high internal consistency ($\alpha = .89$, $\omega = .81$), while extrinsic motivation demonstrated low reliability ($\alpha = .39$, $\omega = .46$), consistent with known limitations of short Work Extrinsic and Intrinsic Motivation subscales (WEIMS) [22]. No reliability was calculated for job satisfaction, as it was measured using a single-item measure.

Construct validity was supported through theoretical grounding and pretesting of newly adapted items. Nevertheless, shortened scales, particularly those with few items, may suffer from reduced content validity, indicating the need for refinement in future research [22][23].

E. Statistical Analyses

To address our research questions, we applied a set of complementary statistical procedures.

First, we used an independent samples t-test to analyze whether the AI use differs by gender. We additionally conducted exploratory group comparisons of AI use across age categories and work-experience categories using one-way ANOVA.

To investigate differences across multiple occupational groups and industries, we conducted a one-way analysis of variance (ANOVA). This analysis assessed whether intrinsic and extrinsic motivation varied depending on participants' professional context (e.g., IT vs. healthcare vs. education) while accounting for work-related AI use. When the omnibus test indicated significant variance across groups, post hoc comparisons were carried out to determine which specific groups differed from one another. This step was essential for revealing patterns that may not be visible in pairwise or aggregated comparisons.

Next, we performed a Pearson correlation analysis to examine the strength and direction of linear associations between key constructs. This included relationships between frequency of AI use and autonomy, between AI use and motivation, and between personality traits (e.g., openness to experience) and perceived job security. Additionally, correlations examined the association between AI use and overall job satisfaction, in line with RQ2. Correlations provided a foundational understanding of how variables co-vary, supporting the identification of potential mediators and predictors relevant to AI acceptance and workplace experiences.

To model the combined influence of multiple predictors, we conducted multiple linear regression analyses. This technique allowed for the simultaneous estimation of the effects of personality traits on AI use, perceived AI-related threat, and perceived autonomy. Regression analyses are especially valuable for isolating unique predictor contributions while controlling for potential confounding variables such as age, experience, or occupational background. Furthermore, the analyses offered insight into whether traits such as openness or neuroticism amplified or attenuated the psychological effects of AI use in the workplace.

Across all statistical analyses, we set significance at $\alpha = .05$. Effect sizes were calculated using Cohen's d for t-tests and η^2 for ANOVA models, interpreted according to Cohen's conventions for small, medium, and large effects to provide meaningful evaluations of practical relevance beyond statistical significance alone [24]. Standard diagnostic tests were conducted to verify model assumptions: Shapiro–Wilk tests were used to assess normality of residuals in regression models, while the Levene test examined homogeneity of variances in ANOVA. These assumptions were largely met, supporting the robustness and validity of the analyses performed.

IV. Results

This section reports the empirical findings of the study, beginning with descriptive results and exploratory analyses of AI use, followed by inferential results concerning autonomy, job

security, motivation, occupational differences, and personality-based predictors of AI use and AI-related perceptions.

A. Descriptive Results

The final sample consisted of 185 participants, of whom 52.4% identified as female and 47.6% as male. The average age was 38.4 years ($SD = 14.2$). Regarding professional experience, 39% reported having more than 20 years of work experience, indicating substantial heterogeneity in career stages.

The distribution across industries showed that the public sector and education were strongly represented with 32.4% ($n = 60$), followed by healthcare and the pharmaceutical industry with 15.7% ($n = 29$). In terms of occupational roles, the largest group consisted of employees (60%, $n = 111$), alongside civil servants (11.4%) and managerial staff (9.7%). This diverse composition enabled differentiated analyses across work contexts.

AI use showed a moderate mean level ($M = 2.75$, $SD = 1.25$), suggesting that AI has been adopted by many employees but is not yet used intensively across all roles. Perceived autonomy ($M = 3.40$, $SD = 0.85$) and perceived job security ($M = 3.15$, $SD = 1.05$) were both rated as moderately high, indicating that employees generally perceived their work environment as stable and supportive despite ongoing technological changes.

B. Exploratory Analysis of AI Use

The exploratory analysis examined the frequency with which employees used AI and whether this usage differed across various demographic factors. Overall usage levels were moderate ($M = 2.75$, $SD = 1.25$), indicating that AI is present in many workplaces but has not yet become ubiquitous or deeply integrated across all job roles.

Employees with less than two years of work experience reported the highest AI use ($M = 3.89$), while those with 11 to 15 years of experience reported substantially lower levels ($M = 2.45$). Descriptively, early-career participants reported higher levels of AI use; this trend may reflect greater openness to experimenting with new technologies and should be examined in future research.

No significant gender differences emerged in AI use ($t = 0.99$, $p = .323$). Similarly, AI use did not differ significantly across age categories ($F(5, 179) = 1.20$, $p = .313$), suggesting that experience and organizational exposure may be more critical than demographic characteristics.

C. Associations between AI Use and Perceived Job Security and Autonomy

To address RQ1, we conducted a bivariate Pearson correlation analysis. Results revealed that AI use showed a significant positive association with perceived autonomy ($r = .24$, $p = .001$) with a small-to-medium effect size. Employees who reported more frequent AI use also reported higher perceived autonomy. However, because the autonomy indicator showed low internal consistency in this sample, this association should be interpreted as preliminary rather than as strong evidence for autonomy-enhancing effects of AI.

In contrast, AI use was not significantly associated with perceived job security ($r = .11$, $p = .153$), indicating a negligible effect size. This indicates that employees did not perceive AI as directly threatening their employment stability. Job security perceptions may be more strongly shaped by organizational communication, economic conditions, or leadership practices than by personal AI use alone.

D. Changes in Motivation and Job Satisfaction

For RQ2, Pearson correlations demonstrated a significant positive relationship between AI use and intrinsic motivation ($r = .23$, $p = .001$), representing a small-to-medium effect size. Employees who interacted more frequently with AI reported greater enjoyment, engagement, and perceived meaningfulness in their work tasks.

AI use showed a positive and statistically significant association with extrinsic motivation ($r = .15$, $p = .042$), indicating a small effect size, although the effect was weaker than for intrinsic motivation. Given the low internal consistency of the extrinsic motivation subscale ($\alpha = .39$), this finding should be treated as tentative. This pattern indicates a small association between AI use and extrinsic motivation, although the estimate should be interpreted cautiously due to measurement limitations.

No significant association emerged between AI use and overall job satisfaction ($r = .08$, $p = .285$), corresponding to a negligible effect size. This indicates that AI use alone does not substantially influence employees' general satisfaction, which may be more strongly tied to broader contextual factors such as leadership, workload, and career opportunities.

E. Differences Across Occupational Groups and Industries

To examine whether motivational effects of AI differed across industries or occupational groups (RQ3), we conducted a one-way ANOVA. No significant differences were found for either intrinsic or extrinsic motivation between industries ($F(7, 177) = 1.378$, $p = .217$) or occupational groups ($F(7, 177) = 0.799$, $p = .589$), and all effect sizes were small ($\eta^2 < .05$).

These results indicate that intrinsic and extrinsic motivation did not significantly differ across work contexts. This aligns with core assumptions of SDT and UTAUT, which posit that perceived autonomy and competence are universal needs that transcend occupational boundaries.

F. Personality Traits as Predictors of AI Use and Perception

To address RQ4, we conducted multiple regression analyses to determine whether personality traits influenced AI use and AI-related perceptions. The results are summarized in Table I

Openness to experience emerged as a significant predictor of AI use ($b = .048$, $p < .001$). Employees high in openness were more likely to adopt AI tools, consistent with research linking openness to curiosity, creativity, and willingness to experiment with new technologies.

Perceived AI-related threat was significantly predicted by neuroticism ($b = .176$, $p = .035$) and extraversion ($b = .166$,

$p = .034$). Individuals high in neuroticism may be more sensitive to uncertainty and risk, while extraverted employees may feel more challenged by changes affecting interpersonal dynamics or work identity.

Perceived autonomy in the context of AI-supported work was significantly predicted by conscientiousness ($b = .017$, $p = .010$). Conscientious employees may perceive AI as a helpful tool for structuring tasks, enhancing efficiency, and supporting goal achievement.

TABLE I. PERSONALITY TRAITS AS PREDICTORS OF AI USE, AI-RELATED THREAT AND AUTONOMY

Predictor (Trait)	AI Use	AI-Related Threat	Autonomy
Openness	$b = 0.048^{***}$	$b = -0.100$	$b = 0.080$
Neuroticism	$b = 0.180$	$b = 0.176^{*}$	$b = 0.070$
Extraversion	$b = -0.160$	$b = 0.166^{*}$	$b = -0.008$
Conscientiousness	$b = 0.003$	$b = 0.015$	$b = 0.017^{*}$
Agreeableness	$b = 0.220$	$b = 0.014$	$b = -0.001$

Significance levels: * $p < .05$, ** $p < .01$, *** $p < .001$.

Overall, these findings highlight the importance of personality traits in shaping employees' perceptions of AI as either a resource or a potential threat.

G. Summary of Key Findings

Across our analyses, AI use was positively associated with perceived autonomy, intrinsic motivation, and, to a lesser extent, extrinsic motivation, while no significant associations emerged for job security or job satisfaction. In addition, openness to experience predicted higher AI use, neuroticism and extraversion predicted greater perceived AI-related threat, and conscientiousness predicted higher perceived autonomy in AI-supported work. Figure 1 provides an integrated overview of the significant associations and predictors identified in Sections

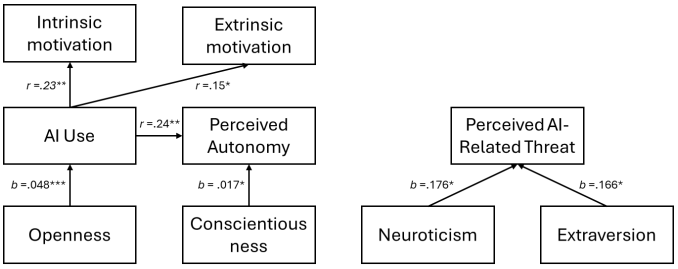


Figure 1. Overview of Significant Associations and Predictors

V. DISCUSSION

This section interprets the results in light of prior research, discusses their implications for organizations and work design, and reflects on the main limitations of the present study.

A. Positioning the Findings Within Existing Research

Our study adds to research on the psychological implications of AI use in the workplace. The positive association between AI use and perceived autonomy aligns with findings that supportive technologies can enhance employees' sense of control [13] and with SDT, which conceptualizes autonomy as a basic psychological need supported when technologies expand individuals' scope of action [5]. Likewise, the link between AI use and intrinsic motivation is consistent with work showing that technology can foster intrinsic motivation by enabling learning opportunities, competence experiences, or creative problem-solving [5][14].

By contrast, AI use was unrelated to perceived job security and overall job satisfaction, deviating from studies that emphasize threat perceptions in the context of automation and digital transformation [8][9]. A plausible explanation is that employees in this sample primarily perceived AI as a complementary tool rather than a substitute for human labour, or that organizational communication and prior digital experience buffered concerns about displacement. These mechanisms were not measured and should be examined in future research; the null finding for job satisfaction also suggests that global satisfaction is more strongly shaped by stable organizational factors than by specific technologies.

The personality findings further highlight individual differences: openness to experience predicted higher AI use [7][18], neuroticism and extraversion predicted higher perceived AI-related threat, and conscientiousness predicted higher autonomy in AI-supported work. Overall, the results support models that integrate technological and psychological dimensions and indicate that AI's impact is contingent on individual dispositions and contextual factors rather than being uniformly positive or negative.

B. Implications for Organizations and Work Design

Our findings suggest that AI can enrich rather than diminish work experiences when it is implemented to support autonomy and intrinsic motivation. AI systems that automate routine tasks or provide decision support can expand employees' decision latitude and foster a stronger sense of control and engagement.

The lack of a negative link between AI use and job security indicates that employees do not automatically view AI as a threat. This offers organizations an opportunity: by emphasizing the augmentative role of AI, communicating transparently about expected changes, and providing training, they can foster trust, acceptance, and psychological safety.

Personality differences further imply that reactions to AI are not uniform. Openness, neuroticism, extraversion, and conscientiousness shape responses to AI, so communication and support should be tailored—for example, by involving more open employees as early adopters and offering additional guidance to those with greater concerns.

Finally, the consistent motivational effects across industries and occupational groups suggest that organizations do not need highly sector-specific strategies. General principles such as autonomy support, transparency, explainability, and employee involvement appear broadly useful and can promote both employee well-being and effective AI implementation.

C. Limitations

This study has several limitations. First, the sample size of 185 participants, although adequate for the analyses, limits

generalizability, especially given the uneven distribution across industries. Public sector and education were overrepresented, whereas technology-intensive and industrial sectors were underrepresented, which may restrict applicability to contexts where AI is used differently or plays a more central operational role.

Second, the cross-sectional design does not allow causal inferences. Although associations between AI use, autonomy, and motivation were found, it remains unclear whether AI use promotes autonomy and intrinsic motivation or whether already autonomous and motivated individuals are more likely to use AI. Longitudinal or experimental designs are needed to clarify directionality and the stability of effects over time.

Third, measurement quality is limited, particularly for the autonomy scale, which showed low internal consistency. In addition, the extrinsic motivation subscale showed low internal consistency, further limiting confidence in the corresponding association with AI use. This suggests that autonomy was not captured with sufficient breadth or precision and that future work should employ more comprehensive or multidimensional measures. The exclusive reliance on self-report also introduces potential response biases (e.g., social desirability, common-method variance) that may have influenced responses and inflated correlations.

Finally, contextual factors that likely shape employees' perceptions of AI were not considered. Variables such as organizational culture, leadership, digital maturity, and communication strategies may be decisive for whether AI is seen as a resource or a threat [16]. Without these influences, the explanatory power of the models is limited, and the interaction between technological and organizational conditions cannot be fully understood. Moreover, the sample was recruited via online networks and mailing lists, which may introduce self-selection bias and limit representativeness

These limitations call for cautious interpretation of the findings and highlight the need for more comprehensive, multi-method, and contextually sensitive research in the future.

VI. Conclusion and Future Work

In our study, we examined the psychological effects of AI use in the workplace, focusing on perceived autonomy, motivation, and job security. We further explored the predictive role of personality traits. The results indicate that higher levels of AI use are associated with stronger perceptions of autonomy and increased intrinsic motivation. These findings support theoretical assumptions from SDT, which argues that technologies enhancing autonomy can foster more meaningful and self-directed work experiences [5]. In this context, more frequent AI use was associated with higher intrinsic motivation and higher scores on an autonomy indicator. However, the autonomy measure's low internal consistency and the cross-sectional design preclude strong conclusions about autonomy-enhancing effects.

In contrast, AI use showed no significant association with perceived job security. This diverges from prior research, which has frequently highlighted automation-related job concerns [8]. The present findings suggest that employees in this sample did not predominantly interpret AI as a threat to employment but rather viewed it as a complementary tool. This divergence underscores the importance of contextual and organizational factors—such as communication, leadership, and prior exposure to digital technologies—that may buffer insecurity and shape how AI is interpreted in practice.

The results concerning personality traits provide additional explanatory insight. Openness to experience emerged as a strong predictor of AI use, supporting the view that open and curious individuals engage more readily with novel technologies. Conversely, neuroticism and extraversion were associated with heightened perceived AI-related threat, indicating that emotional sensitivity and social orientation may amplify concerns in the face of technological change. Conscientiousness was linked to higher perceived autonomy when using AI, suggesting that structured and goal-oriented individuals may integrate AI into their workflow more effectively and view it as a means of improving task management.

Overall, the study demonstrates that the impact of AI on employees is nuanced and shaped by both technological features and individual differences. AI is neither inherently motivating nor inherently threatening; rather, its psychological effects depend on how it is implemented and how employees interpret it. These insights highlight the importance of designing AI-enabled work environments that are transparent, autonomy-supportive, and responsive to diverse employee needs.

Future research should employ longitudinal and experimental designs to examine whether the autonomy- and motivation-related effects of AI use persist over time and whether AI use shapes stable or context-sensitive technology attitudes. Sector-specific studies are needed to clarify how digital infrastructures, job demands, organizational maturity, technological readiness, and task complexity influence employees' responses to AI. Building on the role of personality traits, intervention studies should test whether more personalised training and communication can reduce perceived AI-related threat, particularly among individuals high in neuroticism. Methodological refinements, including improved autonomy measures, data sources beyond self-report, more heterogeneous industry samples, and qualitative approaches, would further strengthen the evidence base. Overall, a nuanced understanding of how AI interacts with motivation, autonomy, and personality will be essential for designing work environments that support well-being, acceptance, and sustainable technological integration.

Ethical Impact Statement

Participation was voluntary and based on informed consent. The survey was conducted anonymously, and no directly identifying personal data was collected. Participants could discontinue participation at any time without providing a reason. Data were stored securely and used exclusively for research purposes in accordance with applicable data protection regulations.

REFERENCES

[1] L. Adolph and A. Tausch, "Künstliche Intelligenz in der Arbeitswelt [Artificial Intelligence in the World of Work]", in *Digitale Arbeit gestalten*, E. Bamberg, A. Ducki, and M. Janneck, Eds., Wiesbaden: Springer Fachmedien Wiesbaden, 2022, pp. 33–47, ISBN: 978-3-658-34646-1 978-3-658-34647-8. DOI: 10.1007/978-3-658-34647-8_3.

[2] Eanes, Z., "IBM pauses hiring on roles AI could replace", Axios Raleigh, Tech. Rep., May 2023.

[3] Klarna, "Klarna AI assistant handles two-thirds of customer service chats in its first month", Tech. Rep., Feb. 2024.

[4] G. M. Spreitzer, "Psychological Empowerment in the Workplace: Dimensions, Measurement, and Validation", *Academy of Management Journal*, vol. 38, no. 5, pp. 1442–1465, Oct. 1995, ISSN: 0001-4273, 1948-0989. DOI: 10.2307/256865.

[5] R. M. Ryan and E. L. Deci, "Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being.", *American Psychologist*, vol. 55, no. 1, pp. 68–78, 2000, ISSN: 1935-990X, 0003-066X. DOI: 10.1037/0003-066X.55.1.68.

[6] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User Acceptance of Information Technology: Toward a Unified View", *MIS Quarterly*, vol. 27, no. 3, p. 425, 2003, ISSN: 02767783. DOI: 10.2307/30036540.

[7] R. R. McCrae and O. P. John, "An Introduction to the Five-Factor Model and Its Applications", *Journal of Personality*, vol. 60, no. 2, pp. 175–215, Jun. 1992, ISSN: 0022-3506, 1467-6494. DOI: 10.1111/j.1467-6494.1992.tb00970.x.

[8] C. B. Frey and M. A. Osborne, "The future of employment: How susceptible are jobs to computerisation?", *Technological Forecasting and Social Change*, vol. 114, pp. 254–280, Jan. 2017, ISSN: 00401625. DOI: 10.1016/j.techfore.2016.08.019.

[9] Eurostat, "Use of artificial intelligence in enterprises", Statistical Office of the European Union (Eurostat), Luxemburg, Tech. Rep., 2024.

[10] A. Milanez, A. Lemmens, and C. Ruggiu, "Algorithmic management in the workplace: New evidence from an OECD employer survey", OECD Artificial Intelligence Papers, Feb. 2025, Edition: 31. DOI: 10.1787/287c13c4-en.

[11] J. Phang et al., *Investigating affective use and emotional well-being on chatgpt*, 2025. arXiv: 2504.03888 [cs.HC].

[12] F. Herzberg, B. Mausner, and B. B. Snyderman, *The motivation to work*, eng, 2. ed. New York: Wiley, 1959, ISBN: 978-0-471-37389-6.

[13] C. Faas, R. Bergs, S. Sterz, M. Langer, and A. M. Feit, *Give Me a Choice: The Consequences of Restricting Choices Through AI-Support for Perceived Autonomy, Motivational Variables, and Decision Performance*, arXiv:2410.07728 [cs], Oct. 2024. DOI: 10.48550/arXiv.2410.07728.

[14] F. Becker, *Mitarbeiter wirksam motivieren: Mitarbeitermotivation mit der Macht der Psychologie [Effectively motivating employees: employee motivation with the power of psychology]*, de. Berlin, Heidelberg: Springer Berlin Heidelberg, 2019, ISBN: 978-3-662-57837-7 978-3-662-57838-4. DOI: 10.1007/978-3-662-57838-4.

[15] P. Berger and J. Von Garrel, "Diffusion of AI value-driven services in the German manufacturing industries—an empirical examination of value-driven service references classified by the business Model Canvas", *Frontiers in Industrial Engineering*, vol. 2, p. 1 407 367, Jul. 2024, ISSN: 2813-6047. DOI: 10.3389/fieng.2024.1407367.

[16] M. Soulami, S. Benchekroun, and A. Galiulina, "Exploring how AI adoption in the workplace affects employees: A bibliometric and systematic review", *Frontiers in Artificial Intelligence*, vol. 7, p. 1 473 872, Nov. 2024, ISSN: 2624-8212. DOI: 10.3389/frai.2024.1473872.

[17] T. A. Judge, D. Heller, and M. K. Mount, "Five-factor model of personality and job satisfaction: A meta-analysis.", *Journal of Applied Psychology*, vol. 87, no. 3, pp. 530–541, Jun. 2002, ISSN: 1939-1854, 0021-9010. DOI: 10.1037/0021-9010.87.3.530.

[18] McElroy, Hendrickson, Townsend, and DeMarie, "Dispositional Factors in Internet Use: Personality versus Cognitive Style", *MIS Quarterly*, vol. 31, no. 4, pp. 809–820, 2007, ISSN: 02767783. DOI: 10.2307/25148821.

[19] A. Unland, M.-A. Heidelmann, and K. Schaaff, "The Influence of Extraversion and Neuroticism on Technology Acceptance and Cognitive Dissonance in LLM Usage", in *COGNITIVE 2026, The Eighteenth International Conference on Advanced Cognitive Technologies and Applications*, 2026.

[20] J. A. Priebe, "Wie man gute von schlechten Tests unterscheidet [how to distinguish good tests from bad ones]", in *Testkonstruktion – das Herz der psychologischen Diagnostik*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2024, pp. 11–32, ISBN: 978-3-662-67546-5 978-3-662-67547-2. DOI: 10.1007/978-3-662-67547-2_3.

[21] H. Moosbrugger and A. Kelava, "Qualitätsanforderungen an Tests und Fragebogen ("Gütekriterien") [quality requirements for tests and questionnaires ("quality criteria")]", in *Testtheorie und Fragebogenkonstruktion*, H. Moosbrugger and A. Kelava, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2020, pp. 13–38, ISBN: 978-3-662-61531-7 978-3-662-61532-4. DOI: 10.1007/978-3-662-61532-4_2.

[22] Tremblay, M. A.; Blanchard, C. M.; Taylor, S.; Pelletier, L. G.; Villeneuve, M., "Work Extrinsic and Intrinsic Motivation Scale: Its value for organizational psychology research", *Canadian Journal of Behavioural Science*, vol. 41, no. 4, pp. 213–226, 2009. DOI: 10.1037/a0015167.

[23] S. D. Gosling, P. J. Rentfrow, and W. B. Swann, "A very brief measure of the Big-Five personality domains", *Journal of Research in Personality*, vol. 37, no. 6, pp. 504–528, Dec. 2003, ISSN: 00926566. DOI: 10.1016/S0092-6566(03)00046-1.

[24] J. Cohen, *Statistical power analysis for the behavioral sciences*, 2nd ed. Hillsdale, N.J: L. Erlbaum Associates, 1988, ISBN: 978-0-8058-0283-2.

A Novel Trap Jamming Technique to Defeat Cognitive Radar

Heath Couture
Mechanical Engineering
University of Waterloo
Waterloo, Canada
e-mail: hcouture@uwaterloo.ca

Qinghan Xiao
Radar Electronic Warfare Section
Defence R&D Canada – Ottawa Research Centre
Ottawa, Canada
e-mail: qinghan.xiao@forces.gc.ca

***Abstract*—Although the cognitive era is still in its infancy, there has been a growing research interest in the development of cognitive capabilities in various electronic systems, such as, cognitive radio and cognitive radar. Historically, there has been a back-and-forth competition between radar and electronic warfare. Thus, with the advances in radar technology, especially the new cognitive radar systems, electronic countermeasures need to be developed to catch up in the race. To defend against cognitive radar systems, a trap jamming technique is proposed to periodically disrupt cognitive radar's measurement capabilities. The algorithm has been developed in a MATLAB environment. The experimental results have shown that the cognitive radar will be rendered ineffective.**

Keywords-jamming; cognitive radar; MATLAB environment.

I. INTRODUCTION

In recent years, there has been a growing research interest in the development of cognitive capabilities in various electronic systems, specifically in the fields of cognitive radio and cognitive radar [1]. The idea of cognitive radio was proposed by Joseph Mitola III in a seminar held at the Stockhom-based KTH Royal Institute of Technology in 1998 [2]. The objective is to enhance the spectrum utilization efficiency by intelligently detecting spectrum holes and rapidly jumping to broadcast on them [3][4]. The concept of cognitive radar was proposed by Haykin in 2006, which presented a dynamic system to adapt and optimize transmitted waveforms based on the operational environment [5]. It was indicated that "Cognitive radar is the radar counterpart to cognitive radio" [6]. With the advances of radar technology, electronic warfare (EW) technology has also improved to disrupt the functionality of radar systems. Radar and EW are always in competition with each other. Therefore, it is necessary to develop a countermeasure technique to defend against cognitive radar systems.

Electronic Attack (EA) or jamming is a key component of EW, and an effective Radio Frequency (RF) technology used to defeat radar systems. Although there are various jamming techniques, they have a common objective — to prevent the proper operation of adversary radars. Since cognitive radars have strong anti-jamming and target detection capabilities, the conventional jamming techniques are unable to interfere effectively with the operation of cognitive radar systems. Based on the principles that "cognitive radar uses the under-utilized spectrum using dynamic spectrum allocation techniques" [7], and "the radar environment is modelled as a Markov decision process to predict the frequency band with the lowest jamming energy [8]", a trap jamming technique is proposed, which will periodically disrupt cognitive radar's measurement capabilities. The algorithms are developed using MATLAB (Version 2023b). The experimental results showed that the proposed technique could disrupt cognitive radar and render it ineffective. The rest of this paper is organized as follows. Section II discusses the different EA techniques. Section III presents a trap jamming approach against cognitive radar. Section IV addresses the development of MATLAB application, while the simulation results are presented in Section V. Finally, Section VI concludes the paper.

II. ELECTRONIC ATTACK TECHNIQUES

Radar is a system that uses electromagnetic waves to identify the range, altitude, direction or speed of objects. It is one of the most powerful and commonly used sensors in the battlefield to detect and track targets such as, aircraft, ships, and vehicles. In contrast, EA, colloquially referred to as jamming, is the electronic countermeasure used to create interference signals to saturate or deceive adversary radars. Different jamming techniques can be employed against radars with a common objective of preventing the proper operation of the radar systems. For example, when detected by a radar, a targeted platform will conduct EA activities to deny range and position information to make the radar lose the tracking information.

A. Passive EA

Passive EA takes place by means of reflection or re-reflecting RF wave energy back to the source to produce false target returns to the radar. Devices used in passive EA include chaff, decoys and other reflectors that require no prime power [9][10].

1) Chaff

Chaff is made of aluminum strip or aluminum-coated nylon or fiber glass, which is developed to create a cloud of false returns on adversary radar systems. It consists of a large number of dipole reflectors that are designed to match the half wavelength of the frequency used by the victim's radar. Chaff with different lengths can be packed in the same package to be effective against radars of widely different frequencies.

2) Passive decoy

Radar corner reflectors are an effective passive decoy against radar detection, which are used to re-radiate relatively

 ISBN: 978-1-68558-375-0

high radar energy mostly back toward the source. Floating corner reflector and corner-reflector decoy were introduced in [11]. With rapid deployment and inflation time, full radar cross-section can be achieved within seconds of reflector launch.

B. Active EA

Electronic jamming is a conventional method of EW, which transmits interfering signals, such as, noise signals or false information, to saturate or deceive the receiver of any electronic device within the range of interference. There are two main techniques of electric jamming: noise techniques and deception techniques [12], while noise jamming can be further categorized as barrage jamming, spot jamming, and sweep jamming.

1) Noise jamming

The objective of noise jamming is to mask the actual signal by introducing an interference signal into the adversary's electronic system. Gaussian noise is the most common noise-jamming waveform. In general, there are three types of techniques for generating the noise signal: wideband jamming, narrow-band jamming, and shifted narrow-band jamming.

a) Barrage jamming

Barrage jamming refers to the use of a single jammer to cover several frequencies with a wide noise bandwidth. In this type of jamming, the power output of the jammer is spread over a bandwidth that is wider than that of several radar signals (Figure 1 (a)). The advantage is that multiple radars can be jammed continuously and simultaneously. The disadvantage lies in that the jammer needs to spread its power over a very wide band, which reduces the jamming power in any one particular band. The wider the frequency band covered, the less effectively each band is jammed.

b) Spot jamming

Spot jamming is simply narrowing the bandwidth of the noise jammer, ideally identical to that of threat emitter frequency, to concentrate the maximum amount of jamming power (Figure 1 (b)). The advantage of spot jamming lies in the jamming efficiency by focusing jammer power on a particular frequency. The disadvantage is its inability to be effective against modern frequency agile radar because the jammer can only jam one frequency. Spot jamming is usually directed against a specific radar. Multiple jammers are required to overcome uncertain frequency parameters.

c) Sweep jamming

To jam a broad band with less output power, sweep jamming repeatedly shifts its full power from one frequency to another (Figure 1 (c)). A group of jamming platforms operating in cooperation may present this scenario [13]. The advantage is that it is able to jam multiple frequencies while maintaining adequate power in quick succession. The disadvantage of sweep jamming is that the jammer cannot affect all the frequencies at the same time.

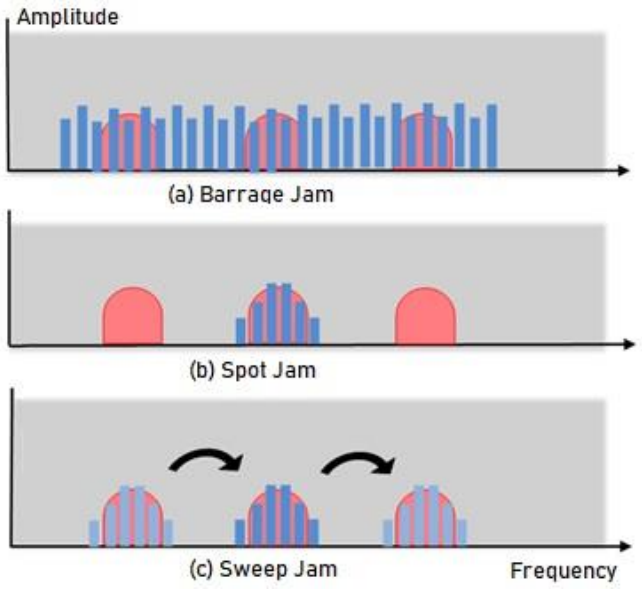


Figure 1. Visual representations of jamming techniques.

2) Deception jamming

Deception jamming consists of manipulating the return signal that the radar receives to generate incorrect data. This can either be through creating false targets or creating misleading range gate information by exploiting Doppler returns. Deception jamming is generally more effective than the noise jamming techniques against modern radars equipped with Electronic Counter-Countermeasures (ECCMs) such as, constant false alarm rate or home-on-jam.

III. TRAP JAMMING TECHNIQUE

As mentioned above, cognitive radar stems from cognitive radio that tries to predict the frequency band with the lowest jamming energy. Therefore, a trap jamming technique is proposed that sets up spectrum holes, lures the cognitive radar jump into one of the spectrum holes, and then generates a jamming signal to cover the spectrum holes. In such a way, the radar cannot lock onto the intended target. The algorithm is explained in detail as follows.

When a radar signal at frequency f_0 is detected using Electronic Support (ES) measures, a low-power noise jamming signal is emitted centred around f_0. Depending on the capacity for the total power output of the EA system, the jam will attempt to create noise coverage of the entire frequency band on which the radar or RF seeker is operating. It can be noted that on larger NATO-designated frequency bands such as, the *J* band (10 to 20 GHz), an RF seeker will only operate on a specific range of the band and not the entire 10 GHz bandwidth. If full band coverage is not possible due to a constraint on the power output of the jammer, noise will be created across a bandwidth from f_1 to f_2 centered around f_0. In this case, the bandwidth will be determined by the emitted power for each Δf that will be high enough to impact the performance of the radar. It is possible that multiple jammers could be used synchronously to cover sufficient bandwidth and meet the power threshold.

 ISBN: 978-1-68558-375-0

As shown in Figure 2, the initial jamming signal transmitted at time t will be similar to a barrage noise jam, with the only difference being that there are two or three spectrum holes of bandwidth b that are not jammed. Using Electronic Intelligence (ELINT) gathered from ES, b should correspond or be close in value to the bandwidth that the radar signal is transmitting. In this case, the Δf of the jamming signal is 10 MHz.

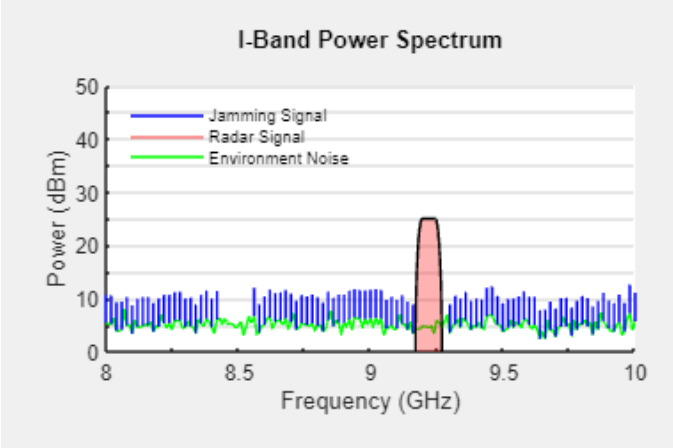


Figure 2. Frequency-power plot at time t, with trap jam signal.

At time t plus an interval of Δt, Figure 3 depicts that the jamming signal will switch to spot jamming with power, P_J, and bandwidth, b, on the "trapped" sections of the frequency domain that were initially left without transmitted noise signals. The Δt intervals will be designated so that the cognitive radar will have ample time to find the trapped locations and start transmitting from them, as shown in Figure 2. The cognitive radar will do this because it will have an optimized Signal-to-Noise ratio (S/N) at the point where there is no interfering noise from the jammer.

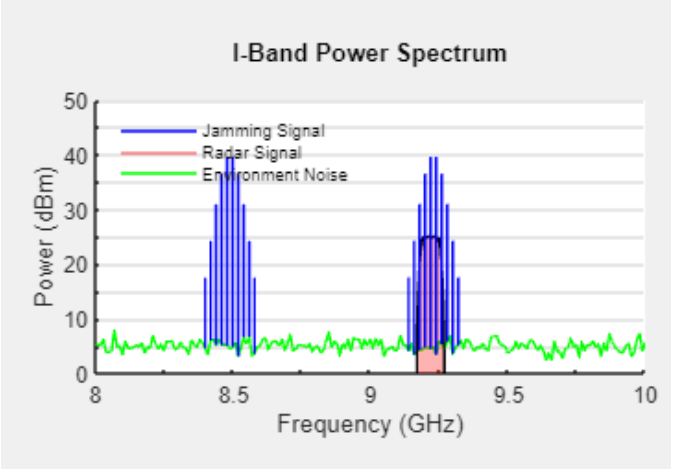


Figure 3. Frequency-power plot at time $t + \Delta t$ with the spot jam signal.

The objective of this Electronic Countermeasure (ECM) is achieved by minimizing the amount of time that the adversary radar is illuminating and locked onto the target. The repetitive nature of this technique will continuously reset adversary radar into its search mode. This also enables the effective use of supplementary ECM. For example, when an RF seeker is in the search mode, launching chaff or deploying a decoy would have a heightened chance of successfully causing the seeker to lock onto those false targets, as opposed to deploying those measures when the seeker is already locked on.

Advantages of this method include no longer needing sophisticated ES measures to gain the exact parameters of the cognitive radar, which could then quickly change regardless. It would be an inefficient use of resources, computational power and time, if every instance the Radar Warning Receiver (RWR) keeps on the illuminated target, and the ESM is employed to identify the radar parameters to generate a spot jam. Therefore, the proposed technique is ideal because ES is only needed for the initial f_0 reading and bandwidth values. After that, the cycles will continue without the delay of an ELINT system to collect and disseminate the parameter values.

IV. MATLAB SIMULATION

The MATLAB simulation of the trap jamming technique was designed with flexibility as the foremost priority, as this is not a comprehensive system design but rather a proof of concept. The development of ECM strategies has additional challenges when only the peacetime operational parameters of radar systems are known in databases with predefined threats. War Reserve Mode (WARM) refers to the use of non-traditional behaviors or modes that are not observed outside of conflicts [14]. This simulation can be used as a guiding tool if those specifications become available.

As previously explored, the cognitive radar allows for the waveform parameters to be flexible, and the simulation incorporates this by making these variables. This MATLAB simulation takes user inputs from the different radar and jammer parameters console, storing them as variables. It then evaluates the signals from both the friendly and adversary sides by using the radar and jammer equations in decibel (dB) form. A Graphical User Interface (GUI) is developed in MATLAB where the user can easily set parameters of the ship, jammer, missile, and radar (Figure 4).

The two-way link radar equation (1) has been standardized in technical literature, and the simulation uses the decibel version (2), which converts the linear parameters to decibel values and includes numerical conversion constants [15]. Using this logarithmic scale allows for comparing high transmitted power values with very low return signal power. This is a standard convention and is used for simplicity and ease of evaluating the simulation results. Signal return power, S, can be written as a function of the range between radar and target. Target detection is also dependent on the S/N ratio, usually in decibels.

 ISBN: 978-1-68558-375-0

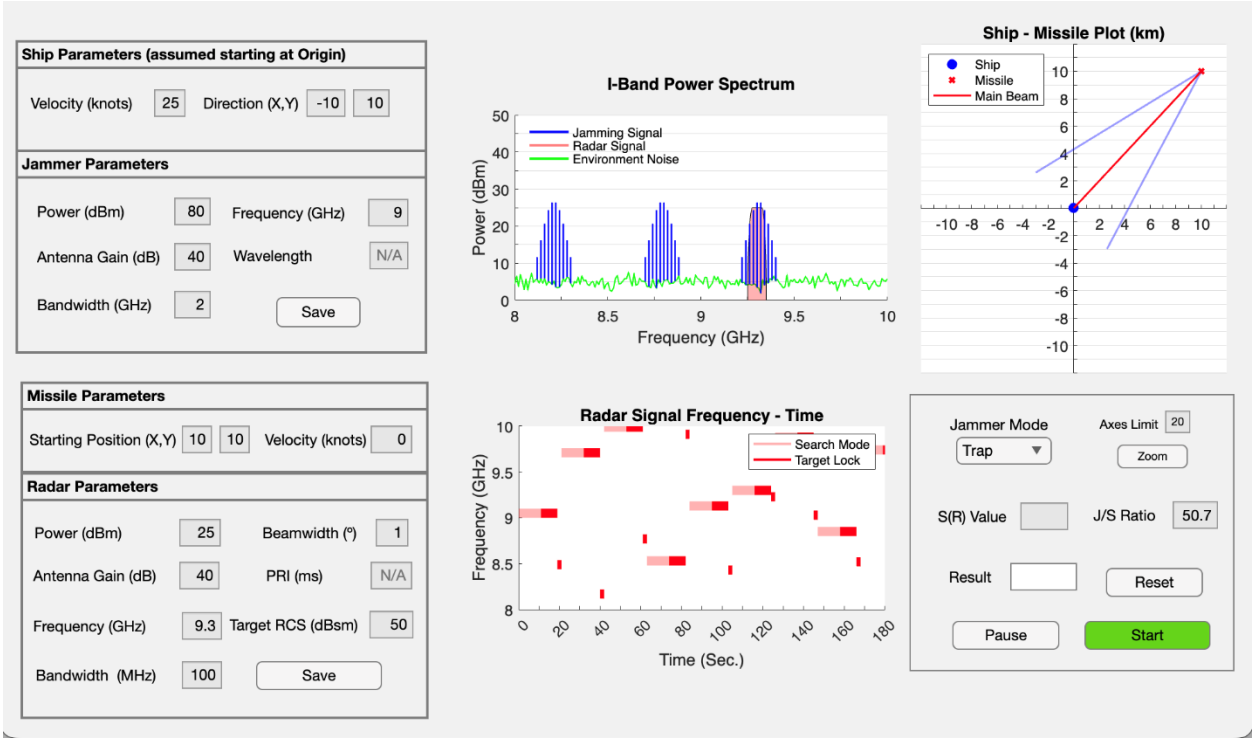


Figure 4. MATLAB GUI.

$$P_r = \frac{P_t G_t G_r \lambda^2 \sigma}{(4\pi)^3 R^4} \quad (1)$$

$$S = P_T + G_T + G_R - 20\log(f) - 40log(R) - 103.4 + \sigma \quad (2)$$

P_r is the return power received by the radar in decibel milliwatts (dBm), P_t is the transmitted power in dBm, G_t is the antenna gain for the transmitter in dB, G_r is the antenna gain for the receiver in dB, λ is the wavelength in meters, σ is the RCS signature of the target in decibel square meters (dBsm), R is the range between the target and radar in km, and f is frequency in MHz.

It is to be expected, and is observed in the simulation with demonstration parameters, that the return power will be a negative number in decibels because this signal power is less than the transmitted power. For simplicity and technical accuracy, it is fair to assume that the same antenna is used for both the transmission and receiving. This is known as a monostatic radar, which uses a transceiver. A missile head will not have the space for multiple antennas. This aspect of the two-way link equation is used more for communications where the transmitter and receiver are not collocated. This restricts the simulation scenarios to the use of Active-Radar Homing (ARH) instead of Semi-Active Radar Homing (SARH), where the missile would only contain a receiver and the transmitted signal would come from an offboard source.

The jamming equation (3) is similar to the radar equation (1), except it is a one-way link. This is because the signal travels directly from the EW device to the receiver of the radar with no return. The jamming signal is a one-way transmission, so it has the advantage of R^{-2} propagation, instead of R^{-4} in (1) [16]. The Jam-to-Signal ratio (J/S) is derived from the decibel form of the radar equation in (2) divided by the decibel version of the one-way link in (4).

$$P_r = \frac{P_j G_j G_r \lambda^2}{4\pi R^2} \quad (3)$$

$$J = P_J + G_J + G_{RJ} - 20\log(f) - 20log(R) - 32.4 \quad (4)$$

$$J/S = P_J + G_J + G_{RJ} - P_T - G_T - G_R + 20log(R) + 71 - \sigma \quad (5)$$

Equation (5) is used as a ratio, and when J/S is less than zero, the radar is effective, but if J/S is greater than zero, the jammer has an advantage and the radar will not operate as effectively [15][17]. It is unknown what the exact threshold of jamming noise is in relation to the signal power that is required to force a cognitive radar to initiate a frequency hop or other parameters modifications. This depends on the specific radar, but for this purpose of proof of concept, the threshold is met when J/S switches signs from positive to negative.

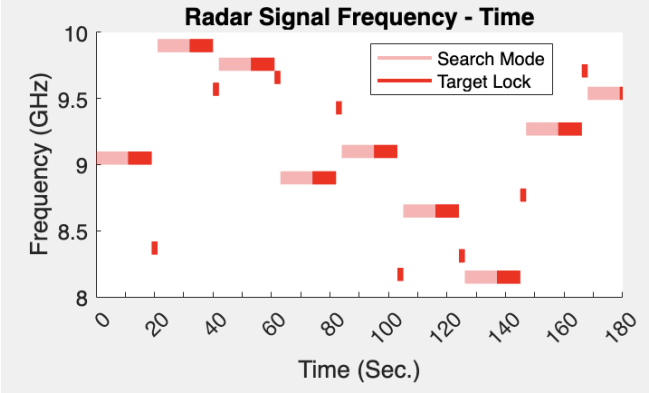
V. SIMULATION EXPERIMENTS

The primary method for evaluation is the radar (2) and jamming (5) equations with test parameters from the simulation. This is then displayed in the numeric edit fields labelled on the GUI. This will give the user confirmation of the mathematically predicted success of the ECM on the cognitive radar. Figure 3 shows that the return pulse is completely hidden. Results with the demonstration parameters

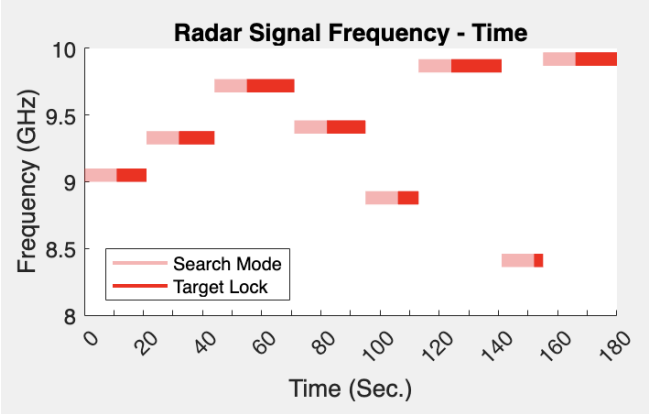
 ISBN: 978-1-68558-375-0

have a J/S ratio of 51.3 when the spot jamming occurs, which confirms this principle.

As aforementioned, the other live-updating figure is displayed in Figure 5. It is a plot of the radar signal frequency as it changes over time, measured in Pulse Repetition Interval (PRI) repetitions. This visualizes how often the cognitive radar is being allured to change its carrier frequency. For target tracking purposes, the jamming method can be considered successful by limiting the amount of time that the radar is locked on or illuminating the intended target.



(a) The radar carrier frequency when the trap jamming method is used



(b) Same plot with the sweep jamming method simulated

Figure 5. Frequency-time plot representing jamming scenario.

There is a drop-down menu that allows the user to select other jamming methods. While no formal comparison exists, although it could and this is a recommendation for improvement, a visual comparison between the time locked on the target for the radar is given if the user runs the simulation multiple times using the trap method, sweep jam, or barrage jam options. Using the Frequency-Time graph shown in Figure 5 (a) to substantiate this, the adversary radar is allured to change transmitting frequency more often when the trap method is used as opposed to sweep jamming, shown in Figure 5 (b), and barrage jamming.

VI. Conclusions and Future Work

A MATLAB interactive simulation of the trap-jamming concept was successfully created to work in all spatial situations and proves the viability of this technique. The desired jamming sequence was achieved through the live updating GUI to replicate the novel trap electronic countermeasure concept (Figure 4). Radar and jamming equation results mathematically prove that the cognitive radar will be rendered ineffective due to the noise transmitted on the radar's operating frequency after it is changed to transmit on the desired, initially unjammed spectrum holes within the operating band. The spot jam will break the lock the radar has which is illuminating the target and reset it back into search mode. While a cognitive radar will learn using reinforcement learning, this concept shows that this cycle will at least be successful once, which is all it might need to be in a real-world EW scenario.

Possible future works include simulation supplemented with improvements such as, implementing Artificial Intelligence (AI) on the radar to have it learn ECCMs, and the addition of other supporting ECM such as, chaff or decoys. The user experience can be enhanced by adding a way to speed up time, from seeing each pulse in microseconds to real-time to see the missile and target moving. These are top areas of future research.

References

[1] W. Hilal, S. A. Gadsden, and J. Yawney, "Cognitive Dynamic Systems: A Review of Theory, Applications, and Recent Advances," in Proceedings of the IEEE, vol. 111, no. 6, pp. 575-622, 2023.

[2] A. Sarode and P. Ojha, "Cognitive Radio," International Research Journal of Innovations in Engineering and Technology (IRJIET), vol. 5, no. 8, pp 71-74, 2021.

[3] S. Pavithra, S. Karthikeyan, V. J. K. Sonti, and S. Jayashri, "Competent Realisation of Cooperative Spectrum Sensing in Cognitive Radio Systems," International Journal of Engineering Systems Modelling and Simulation, vol. 7, no. 2, pp. 103-110, 2015.

[4] K. B. Letaief and W. Zhang, "Cooperative Communications for Cognitive Radio Networks," in Proceedings of the IEEE, vol. 97, no. 5, pp. 878-893, 2009.

[5] S. Haykin. "Cognitive Radar: A Way of The Future," IEEE Signal Processing Magazine, vol. 23, pp. 30-40, 2006.

[6] C. Baylis, J. Martin, M. Moldovan, O. Akinbule, and R. J. Marks, "A Test Platform for Real-Time Waveform and Impedance Optimization in Microwave Radar Systems," 2012 International Waveform Diversity & Design Conference (WDD), Kauai, HI, USA, 2012, pp. 019-022, doi: 10.1109/WDD.2012.7311307.

[7] D. Brahma, S. Swayamsiddha, and G. Panda, "Dynamic Spectrum Allocation in Cognitive Radar: A Brief Overview," 2023 3rd International Conference on Range Technology (ICORT), Chandipur, Balasore, India, pp. 1-4, 2023.

[8] Z. Zheng, W. Li, and K. Zou, "Airborne Radar Anti-Jamming Waveform Design Based on Deep Reinforcement Learning," Sensors 2022, 22 (22): 8689.

[9] A. De Martino, Introduction to Modern EW Systems, Norwood: Artech House, 2012.

 ISBN: 978-1-68558-375-0

[10] M. I. Skolnik, Radar Handbook, 3rd Edition, New York: McGraw-Hill, 2008.

[11] N. Friedman, "Soft Kill Versus Anti-Ship Missiles," Naval Forces, vol. 30, no. 1, pp. 85-89, 2009.

[12] Electronic Warfare And Radar Systems Engineering Handbook, *NAVAIR Electronic Warfare/Combat Systems*, June 2012, [Online]. Available from: http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA566236.

[13] F. Schmied and J. Schobel, "Algorithmic Optimization of Sweep-based Signals for Jamming RF Modules and UAVs," STO-MP-IST-205-37, pp. 1-8, 19, 2024.

[14] P. M. Gale, "Counter ESM/ELINT - A Review," Maple Leaf Chapter Newsletter - Association of Old Crows, vol. 1, no. 7, pp. 3–6, 2023.

[15] D. Adamy, "EW Against Modern Radars - Part 1 Radar Jamming Equations," Journal of Electronic Defense, pp. 56–57, 2009.

[16] M. Davis, "Key Differences between Radar and Communications Systems," in Proceedings of the 12th Annual International Symposium on Advanced Radio Technologies, pp.1-11, 2011.

[17] D. Adamy, EW 101: A First Course in Electronic Warfare. Boston: Artech House, 2001.

 ISBN: 978-1-68558-375-0

A Hybrid Cognitive Architecture for Multimodal and Multilingual Human–Machine Interaction

Nana Schlage
Institute of Engineering and Computer Science
Niederrhein University of Applied Sciences
Krefeld, Germany
nana.schlage@hs-niederrhein.de

Toni Thelen
Institute of Engineering and Computer Science
Niederrhein University of Applied Sciences
Krefeld, Germany
toni.thelen@hs-niederrhein.de

Lukas Cramer
Institute of Engineering and Computer Science
Niederrhein University of Applied Sciences
Krefeld, Germany
lukas.cramer@hs-niederrhein.de

Edwin Naroska
Institute of Engineering and Computer Science
Niederrhein University of Applied Sciences
Krefeld, Germany
edwin.naroska@hs-niederrhein.de

Gudrun Stockmanns
Institute of Engineering and Computer Science
Niederrhein University of Applied Sciences
Krefeld, Germany
gudrun.stockmanns@hs-niederrhein.de

***Abstract*—Public spaces, such as libraries, play a key role in providing equitable access to information. Yet, current digital and robotic services often struggle to address linguistic diversity, multimodal accessibility, and inclusive interaction needs, challenges that require advanced cognitive technologies. This paper presents a modular, service-oriented architecture designed to enable accessible, multilingual, and multimodal interaction in public library environments. The system combines cognitive perception (voice activity detection and multilingual speech transcription), hybrid intent understanding (local transformer-based classification model with Large Language Model-supported reasoning), and a controllable dialogue management mechanism that integrates symbolic dialogue graphs with Large Language Model generation. Knowledge access is realized through a dual recommendation pipeline that merges Retrieval-Augmented Generation with a tool-augmented Large Language Model agent for context-aware information delivery. All components operate as containerized services within a flexible client–server architecture, ensuring hardware independence, maintainability, and local control of all processes. The system emphasizes transparency and safety by constraining generative models through predefined graph structures, enabling predictable behavior and preventing sensitive or inappropriate topics from being addressed while preserving the adaptivity of Large Language Model-based reasoning. Initial deployments in a library environment show promising interaction possibilities in multilingual environments and effective cognitive support, demonstrating the architecture's potential as a trustworthy and inclusive cognitive service platform.**

Keywords-cognitive architecture; multimodal dialogue systems; multilingual interaction; large language models; knowledge retrieval and reasoning.

I. Introduction

Human–Robot Interaction (HRI) in public spaces has gained increasing relevance as institutions, such as libraries, museums, hospitals, and municipal offices, seek to improve equitable access to information and services. Social robots offer potential to provide orientation, information retrieval, and assistance through natural multimodal interaction, lowering access barriers for diverse user groups [1]–[4].

Despite this potential, current HRI systems face persistent limitations. Many rely on monolingual interfaces, rigid rule-based dialogue strategies, or narrowly scoped interaction flows that fail to accommodate heterogeneous user populations, including non-native speakers, elderly individuals, children, and people with disabilities [5][6]. These constraints are exacerbated in real-world deployments, where interaction requirements are unpredictable and must remain robust, transparent, and socially appropriate.

Recent advances in Large Language Models (LLMs) provide capabilities for multilingual communication, intent inference, and context-aware dialogue generation [3][7]. However, challenges, such as response latency, limited controllability, and ethical risks, hinder their direct use in public-facing robots [8][9]. Purely generative dialogue approaches alone are therefore often unsuitable for safety-critical or socially sensitive environments, such as libraries.

Inclusive HRI increasingly emphasizes multimodality, combining speech, visual feedback, and touch-based interfaces to address diverse accessibility needs [10]. Yet, many implementations treat multimodality, multilinguality, and dialogue intelligence as isolated design problems, resulting in fragmented, ad hoc architectures that are difficult to scale, maintain, or generalize.

This paper addresses these challenges by presenting a modular, service-oriented cognitive interaction system for deployment in public libraries. It integrates multilingual speech interaction, multimodal interfaces, and adaptive dialogue capabilities within a unified architecture. Central to the approach is a

 ISBN: 978-1-68558-375-0

hybrid dialogue management mechanism combining predefined symbolic graphs with LLM-based response generation. By constraining generative models through structural guidance, the system achieves predictable, context-appropriate dialogue while preserving adaptive reasoning.

The main contributions of this paper are as follows:

1) **A modular, service-oriented system architecture** for multimodal and multilingual HRI, designed for hardware independence, maintainability, and controlled deployment.
2) **A hybrid dialogue management framework** combining symbolic interaction graphs with LLM-based responses, balancing adaptability with predictable, transparent behavior.
3) **Technical validation through component-level testing**, informed by initial library deployments, demonstrating feasibility and practical benefits for inclusive public interaction.

The remainder of this paper is organized as follows. Section II reviews related work on HRI, dialogue management, and multilingual inclusive interaction. Section III presents the system overview and key components. Section IV outlines the evaluation methodology and results from component-level testing. Section V concludes and outlines directions for future work.

II. Related Work

Social robots have been deployed in libraries, museums, hospitals, and other public spaces, primarily supporting wayfinding, information provision, or engagement. But real-world environments remain challenging due to noise, heterogeneous users, and dynamically changing contexts [1]. Although recent work increasingly integrates conversational Artificial Intelligence (AI), many deployed systems still rely on monolingual, rule-based interactions with limited adaptability and inclusiveness.

Based on these challenges, this paper reviews three research areas: (i) HRI in public spaces, (ii) dialogue management and large language models in social robots, and (iii) multilingual and inclusive interaction.

A. HRI in Public Spaces

Robots, such as Pepper, Sanbot, and Temi, have been used in public environments primarily as guides or information providers [1]. While these studies demonstrate feasibility in public deployments, systems often struggle with robustness and adaptability when faced with heterogeneous users and dynamically changing interaction contexts [11]. Recent work, such as the Navel robot, explores the integration of LLM-based dialogue to enable more natural interaction. However, most existing approaches target semi-controlled environments (e.g., care facilities) and are less suited for highly dynamic, noisy public spaces with varied and anonymous users. These contexts require modular architectures, multimodal support and reliable safeguards, particularly where interactions may involve children or vulnerable groups.

B. Dialogue Management and Large Language Models in Social Robots

Dialogue management remains central to social robots. Traditional rule-based or finite-state systems offer transparency but are limited in flexibility and scalability in open-domain conversations [12]. Data-driven and reinforcement learning approaches address adaptability in dialogue management but introduce challenges related to data requirements, complexity and real-world generalization. Recent surveys highlight these issues and discuss current trends in neural and adaptive dialogue management techniques for HRI [13]. Hybrid architectures increasingly combine symbolic control with machine learning to balance predictability and responsiveness [9].

Recently, LLMs, such as GPT-4 and PaLM have enabled more open-ended, context-aware interaction capabilities by maintaining conversational context over multiple turns and generating coherent responses in complex settings [14]. While systems like Xiaoice demonstrate engaging long-term conversational capability, applying LLMs in embodied agents introduces challenges regarding latency, controllability, safety and bias [8]. Many existing robot implementations therefore remain monolithic and difficult to extend, underlining the need for modular, safe and scalable dialogue solutions for public environments.

C. Multilingual and Inclusive Interaction

Public institutions serve diverse audiences, yet many deployed robots still support only a single language and provide limited accessibility features [15]. Recent work on socially assistive robots integrating large language models demonstrates how multimodal dialogue systems can support multilingual and adaptive interaction in real-world environments, facilitating meaningful engagement with diverse user groups, including older adults and socially isolated individuals [16]. Other work has explored adaptive dialogue for users with cognitive impairments [17]. However, many systems remain constrained to pre-scripted interactions and lack real-time adaptability.

Multimodal interfaces combining speech, visual display and touch can improve inclusiveness and interaction robustness, but are often implemented in an ad hoc manner without forming extensible frameworks [10]. Recent Human-Computer Interaction (HCI) research emphasizes the need to account for linguistic, cognitive and sensory diversity in public-facing technologies [15], yet few robotic systems integrate these principles holistically.

D. Comparison to Existing Systems

Unlike fixed-script or monolithic robots, the proposed system adopts a modular, graph-based dialogue architecture, enabling non-linear flows, multilingual support, and multimodal input while ensuring predictable and safe behavior. Decoupling dialogue logic from hardware enhances portability, maintainability, and scalability compared to prior approaches.

III. System Overview

This paper presents a modular, service-oriented HRI system for personalized, multimodal interaction in public libraries. All core computation and decision-making processes are performed locally on a single host computer, while the robot platform serves solely as an actuator, minimizing dependencies, ensuring

hardware independence, and providing a controlled public-facing environment.

The architecture is designed to balance adaptive cognitive capabilities with strict control requirements typical for socially sensitive public environments. Cognitive reasoning, multilingual understanding, and multimodal interaction are integrated while maintaining predictable and transparent system behavior.

The system's functionality is decomposed into specialized services, each responsible for a stage of the interaction pipeline, enabling independent development, testing, and controlled integration of cognitive components:

- **Speech Detection:** continuously identifies spoken input using silence and Voice Activity Detection (VAD).
- **Speech Processing:** transcribes speech and detects language, supporting robust multilingual interaction.
- **Intent Classification:** infers user intent via Bidirectional Encoder Representations from Transformers (BERT)-based embeddings with GPT-4o fallback, allowing fast inference and robustness in ambiguous cases.
- **Dialogue Management:** controls interaction via a structured graph, combining symbolic transitions with LLM-generated utterances, balancing flexibility with predictable, traceable behavior.
- **Book Recommendation:** generates personalized suggestions through a dual pipeline, Retrieval-Augmented Generation (RAG) for fast responses and a tool-enabled LLM agent for multi-step exploratory searches.
- **Data Server:** coordinates inter-service communication via WebSocket publish/subscribe patterns and logs interactions.
- **Graphical User Interface (GUI):** visualizes system state, dialogue nodes, and transitions, supporting debugging and simulated input injection.

Each service executes in its own Docker container, ensuring isolated execution, reproducibility, and maintainability. The decoupled design allows services to be updated, replaced, or extended independently without affecting overall system behavior, which is critical for long-term deployments in public cognitive systems.

A. Interaction Flow

Figure 1 illustrates the overall interaction flow. Audio input is continuously monitored by the speech detection module, transcribed and language-identified by the speech processing service, and passed to the intent classifier. Recognized intents and the user input are sent to the dialogue manager, which determines responses using the interaction graph, triggering verbal replies, book recommendations, or other actions.

In parallel, the system supports touch-based input through the graphical user interface of the robot. These inputs bypass the speech pipeline and are sent directly to the dialogue manager, ensuring accessibility and robustness in noisy environments or for users with speech impairments.

B. Speech Detection and Processing

Speech detection combines silence-based triggering with neural voice activity analysis to minimize unnecessary computation.

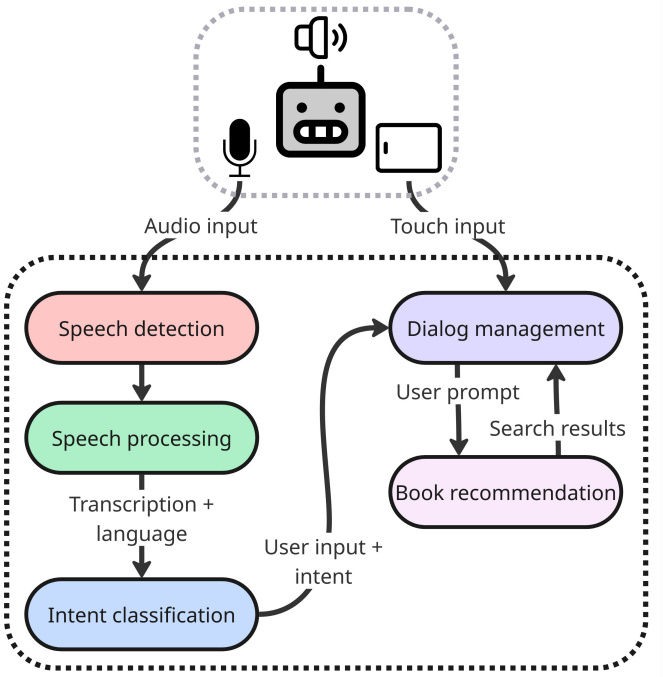


Figure 1. General interaction flow.

First, a lightweight wake-word and silence detector based on Precise [18] identifies candidate segments. These segments are then refined using Silero VAD [19]. Only confirmed speech is forwarded for transcription, reducing both latency and resource usage.

Speech processing is performed using the Whisper Large-v2 model via the WhisperX library [20], enabling multilingual transcription and automatic language identification under real-world conditions.

C. Intent Classification

Intent classification follows a hybrid strategy: a lightweight multilingual BERT-based sentence embedding model (T-Systems-onsite/cross-en-de-roberta-sentence-transformer [21]) compares inputs to predefined intent examples using cosine similarity, enabling fast local inference. In cases of low confidence, GPT-4o serves as a fallback, which takes the interaction context into account.

These newly classified examples are subject to validation and can be reviewed or removed during maintenance, supporting incremental refinement and robust handling of ambiguous cases while mitigating error accumulation.

To support multilingual users, transcriptions are either processed directly in the languages supported by BERT (German or English) or translated into English using the MyMemory Translator API. The translation service is easily replaceable and employed only because tested local alternatives did not achieve sufficient accuracy in source languages Arabic and Turkish.

D. Dialogue Management and Interaction Graph

The dialogue management framework employs a directed interaction graph, where nodes represent dialogue states and edges define transitions based on user intent or system conditions. This structure consists of a *main graph* for high-level interaction states (e.g., start screen, opening hours, events) and *subgraphs* for reusable routines (e.g., language selection, favorites management, multi-step queries, book-specific information requests). The default subgraph is active across all main states, while others are assigned selectively, supporting modularity, reusability, and simplified maintenance.

To enable flexible control, the graph supports intent-, time-, condition-, and flow-based transitions with adjustable priorities, enabling user- and system-driven progression.

Nodes can execute logic on entry or exit, including conditional checks, data updates, and output generation, enabling context-sensitive responses.

Figure 2 shows an example with three nodes, illustrating how intents and example utterances guide transitions between states.

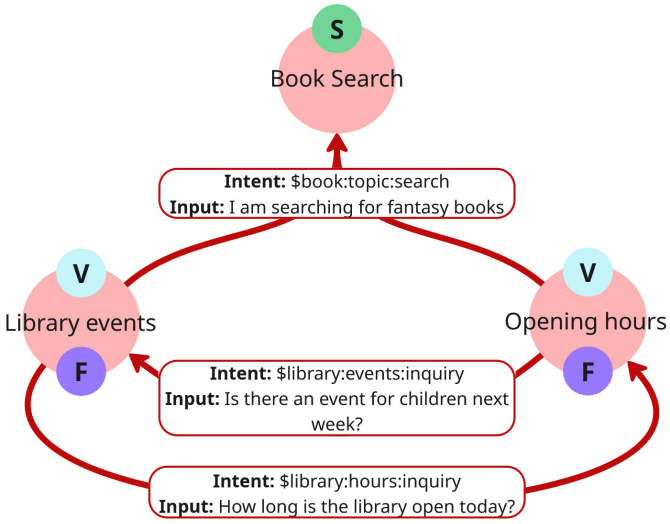


Figure 2. Example dialogue interaction graph with three nodes.

While the interaction graph governs dialogue structure and control flow, selected system utterances are generated dynamically using GPT-4o-mini under node-specific constraints. This hybrid approach combines structured dialogue control with context-aware language generation, maintaining transparency and traceability, key features for cognitive systems in public environments.

The system currently supports German, English, Turkish, and Arabic, and can be extended to additional languages with minimal effort, allowing inclusive interaction across diverse user populations.

E. Graphical User Interface (GUI)

The implemented GUI supports development, monitoring, and debugging by visualizing the current dialogue state, active nodes, and possible transitions. It allows simulated inputs for testing without relying on the robot platform.

F. Book Recommendation

To provide personalized book suggestions, the system combines a RAG pipeline with a tool-enabled LLM-based search agent. The RAG pipeline operates on the internal catalog of 42,000 books for rapid semantic retrieval, ensuring efficient responses. For underspecified or exploratory requests, the LLM agent conducts multi-step searches, drawing not only on the local catalog but also leveraging external sources, thereby covering a broader range of books and user interests. To reduce latency, the same BERT-based embedding model used for intent classification serves as a semantic pre-filter for the LLM agent, ensuring context-aware and responsive recommendations. This dual-pipeline design balances efficiency with flexible, user-tailored exploration, allowing the system to provide both fast and comprehensive suggestions.

IV. Evaluation

Prior to the structured evaluation reported in this section, the system underwent iterative testing in lab experiments, two public libraries, and simulated interactions with a virtual user (GPT-4o-mini). These formative tests identified usability issues, validated core functionalities, and informed architectural refinements. The resulting adapted system is evaluated in this paper.

The evaluation focuses on technical validation of system components and overall interaction behavior under controlled scenarios. No formal user study was conducted. Scenarios assess feasibility, robustness, and practical suitability for inclusive library interactions.

Scripted flows cover common tasks, such as book searches, information requests, and language switching, as well as edge cases like repeated, ambiguous, or off-scope queries to test redundancy handling, off-topic detection, and error recovery. Metrics were recorded automatically to ensure reproducibility.

A. Translation Benchmark

Five translation solutions were evaluated for multilingual interaction: two fully local models (nllb-200-3.3B, Argos Translate) and three API-based services (ChatGPT 4o, Google Translator, and MyMemory Translator). The evaluation considered translation quality, latency, and deployment-related aspects, such as privacy implications and cost.

Table I summarizes the benchmark results and key observations.

TABLE I. Translation test results for different models

Model	ØLatency [s]	Key Findings
nllb-200-3.3B	**0.33**	Unstable translations
Argos	0.66	Unstable translations
ChatGPT 4o	2.01	High quality, costly
Google	1.79	Good quality, free tier
MyMemory	1.65	High quality, free tier

Local models had lowest latency but sometimes produced unintended or mixed-language outputs, rendering them unsuitable for reliable deployment in public-facing systems. Among the API-based approaches, MyMemory Translator provided the best balance between translation quality, reliability, and response time while offering a limited free usage tier. Although this approach involves transmitting text data to an external service, no audio data or user identifiers are shared, and observed latency remained acceptable for interactive use.

Future work may explore optimized local translation models that reduce latency while fully avoiding third-party data transmission.

B. Intent Classification Benchmark

Six models were benchmarked on 372 manually labeled German and English utterances to identify the best intent classifier. Each utterance was assigned a ground-truth intent label, enabling direct comparison of classification accuracy. Average inference time per utterance was measured to account for latency constraints in interactive settings.

Table II summarizes the benchmark results.

TABLE II. COMPARISON OF EVALUATED INTENT CLASSIFICATION MODELS

Model	ØLatency [s]	Error Rate [%]
Infloat E5	0.22	46.77
Infloat E5 (norm.)	**0.06**	41.67
Qwen3	0.15	31.99
Qwen3 (norm.)	0.09	46.77
BGE-M3	0.10	13.44
BERT	0.23	**3.23**

Although Infloat E5 and Qwen3 had lower inference times, their error rates were too high for real-world use. BGE-M3 presented a reasonable compromise between latency and accuracy but was clearly outperformed by the BERT-based model, which achieved the lowest error rate despite slightly higher inference time.

Based on these results, the BERT-based model was selected as the primary intent classifier. To mitigate rare misclassifications, it is complemented by a GPT-4o fallback mechanism, as described in Section III-C.

Future work may investigate domain-adaptive fine-tuning of the BERT-based classifier to further improve intent recognition accuracy and robustness across languages.

C. End-to-End System Evaluation

End-to-end evaluation was conducted using the scripted interaction flows described above, executed in German, English, Turkish, and Arabic. This multilingual setup enabled assessment of combined system behavior, including translation, intent classification, dialogue control, and book recommendation.

Table III summarizes the results.

TABLE III. END-TO-END SYSTEM EVALUATION METRICS (DE: GERMAN, EN: ENGLISH, TR: TURKISH, AR: ARABIC, EV: EVALUATION)

	DE	EN	TR	AR	EV
Test Cases	247	254	253	251	$\sum$ 1005
ØResponse time [s]	2.68	2.68	3.44	5.04	Ø3.46
ØTime book searches [s]	12.1	19.2	15.3	17.8	Ø15.7
Success rate [%]	98.8	97.2	96.4	96.4	Ø97.2
Intent error rate [%]	1.21	1.97	2.37	2.79	Ø2.09
Intent fallbacks	47	51	63	58	$\sum$ 219

Across 1005 scripted test cases, the system achieved a mean interaction success rate of 97.2%, 3.46s average response time, and a mean intent error rate of 2.09%. Turkish and Arabic had slightly higher response times and fallback usage due to translation and intent complexity. Nevertheless, overall performance remained within acceptable bounds for interactive public deployments.

No direct baseline comparison to existing library robot systems was performed, as comparable multilingual and multimodal systems with controlled dialogue constraints are not publicly available. Instead, evaluation focuses on internal consistency and component-level performance under realistic usage scenarios.

Future development should prioritize replacing remaining external services (e.g., translation services) with fully local components wherever feasible. Such an approach would enhance data privacy, reduce dependency on third-party providers, and further improve latency consistency in real-world deployments.

D. Lessons Learned

The evaluation highlights the effectiveness of hybrid architectures that combine local models with LLM-based fallback mechanisms to balance performance, robustness, and adaptability. Structured dialogue control proved essential for maintaining predictable behavior when integrating generative models. At the same time, reliance on external services introduced variability in latency and raised privacy considerations, reinforcing the importance of fully local alternatives for future deployments.

E. Ethical and Privacy Considerations

The use of external generative services raises concerns regarding data privacy, transparency, and ethical accountability. The system therefore processes sensitive data locally whenever possible. Only text strictly required for translation, fallback intent classification, or response generation is transmitted to external services, and no audio data or user identifiers are shared. Interaction logs are anonymized and access is restricted to authorized personnel.

By supporting multilingual and culturally sensitive interaction without requiring user identification, the system promotes fairness, inclusivity, and trust. Future work aims to replace remaining cloud-based services with local alternatives, further strengthening compliance with data protection regulations and ethical standards expected in public-sector applications.

V. CONCLUSION AND FUTURE WORK

This paper presented a modular, service-oriented HRI system enabling multilingual and inclusive interaction with public

library resources. The architecture supports predictable, context-aware interaction across speech and touch modalities through a hybrid intent classification pipeline, a graph-based dialogue manager, and a two-stage book recommendation approach.

Component-level evaluation demonstrated the system's feasibility for public library scenarios. Across 1005 scripted interactions in four languages, the system achieved an overall interaction success rate of 97.2% with a mean intent error rate of 2.09%. GPT-4o fallback was required only rarely, indicating reliable intent recognition and stable dialogue control even under challenging conditions, such as repeated or out-of-scope requests.

The choice of a BERT-based intent classifier reflects a deliberate trade-off favoring robustness over minimal latency. Despite slightly higher inference times, its substantially lower error rate and local execution reduce reliance on cloud-based fallbacks and contribute to predictable response behavior in public interactive settings.

Recent advances in generative AI are likely to further improve translation quality and intent recognition capabilities in the near future. While such progress may reduce certain performance gaps, empirical validation remains essential, particularly in public-sector HRI contexts. Domain-specific requirements, such as privacy preservation, predictable dialogue behavior, latency consistency, and multilingual robustness, cannot be guaranteed by model scaling alone. The hybrid architecture presented in this work therefore reflects a structural design choice rather than a temporary workaround, while remaining adaptable to future technological developments.

Beyond technical performance, the system demonstrates how conversational robots can lower access barriers to public knowledge. Support for German, English, Turkish, and Arabic addresses linguistic diversity and promotes inclusive access in heterogeneous urban communities, reinforcing public libraries as accessible, digitally mediated spaces.

Future work will prioritize replacing remaining external services, particularly translation components, with fully local alternatives to improve privacy and latency consistency. Additional directions include domain-adaptive fine-tuning of the intent classifier, support for additional languages, and the integration of user feedback to enable adaptive and personalized interaction over time. These developments aim to further advance privacy-preserving, scalable conversational agents for deployment in public-sector knowledge environments.

ACKNOWLEDGMENT

The presented work was supported by the RuhrBots competence center (16SV8693) funded by the Federal Ministry of Research, Technology and Space of Germany.

REFERENCES

[1] O. Mubin, I. Kharub, and A. Khan, "Pepper in the library" students' first impressions", in *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, ser. CHI EA '20, Honolulu, HI, USA: Association for Computing Machinery, 2020, pp. 1–9, ISBN: 9781450368193. DOI: 10.1145/3334480.3382979.

[2] L. C. Nguyen, "The impact of humanoid robots on australian public libraries", *Journal of the Australian Library and Information Association*, vol. 69, no. 2, pp. 130–148, Apr. 2020. DOI: 10.1080/24750158.2020.1729515.

[3] M. Rohrmüller *et al.*, "Perspectives on using multi-modal large language models for physical human-robot interaction", ser. ICRA, Sep. 2024.

[4] Y. Lai *et al.*, *Fam-hri: Foundation-model assisted multi-modal human-robot interaction combining gaze and speech*, 2025. arXiv: 2503.16492 [cs.HC].

[5] S. O. Oruma, M. Sánchez-Gordón, R. Colomo-Palacios, V. Gkioulos, and J. K. Hansen, "A systematic review on social robots in public spaces: Threat landscape and attack surface", *Computers*, vol. 11, no. 12, 2022, ISSN: 2073-431X. DOI: 10.3390/computers11120181.

[6] C. Toussaint, P. T. Schwarz, and M. Petermann, "Navel - a social robot with verbal and nonverbal communication skills", in *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, ser. CHI EA '23, Hamburg, Germany: Association for Computing Machinery, 2023, pp. 1–4, ISBN: 9781450394222. DOI: 10.1145/3544549.3583898.

[7] P. Allgeuer, H. Ali, and S. Wermter, "When robots get chatty: Grounding multimodal human-robot conversation and collaboration", in *Artificial Neural Networks and Machine Learning – ICANN 2024*. Springer Nature Switzerland, 2024, pp. 306–321, ISBN: 9783031723414. DOI: 10.1007/978-3-031-72341-4_21.

[8] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big? [parrot]", in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '21, Virtual Event, Canada: Association for Computing Machinery, 2021, pp. 610–623, ISBN: 9781450383097. DOI: 10.1145/3442188.3445922.

[9] J. Wang *et al.*, *Large language models for robotics: Opportunities, challenges, and perspectives*, 2024. arXiv: 2401.04334 [cs.RO].

[10] H. Su *et al.*, "Recent advancements in multimodal human–robot interaction", *Frontiers in Neurorobotics*, vol. 17, 1084000, 2023, ISSN: 1662-5218. DOI: 10.3389/fnbot.2023.1084000.

[11] H. Yoon, G. Shim, H. Lee, M.-G. Kim, and S. Kim, "Observation of human–robot interactions at a science museum: A dual-level analytical approach", *Electronics*, vol. 14, no. 12, p. 2368, 2025. DOI: 10.3390/electronics14122368.

[12] H. Brabra *et al.*, "Dialogue management in conversational systems: A review of approaches, challenges, and opportunities", *IEEE Transactions on Cognitive and Developmental Systems*, vol. 14, no. 3, pp. 783–798, 2022. DOI: 10.1109/TCDS.2021.3086565.

[13] M. M. Reimann, F. A. Kunneman, C. Oertel, and K. V. Hindriks, "A survey on dialogue management in human–robot interaction", *ACM Transactions on Human–Robot Interaction*, vol. 13, no. 2, p. 22, 2024. DOI: 10.1145/3648605.

[14] Y. Kim *et al.*, "A survey on integration of large language models with intelligent robots", *Intelligent Service Robotics*, vol. 17, pp. 1091–1107, 2024. DOI: 10.1007/s11370-024-00550-5.

[15] K. Seaborn, G. Barbareschi, and S. Chandra, "Not only weird but "uncanny"? a systematic review of diversity in human–robot interaction research", *International Journal of Social Robotics*, vol. 15, no. 11, pp. 1841–1870, 2023, ISSN: 1875-4805. DOI: 10.1007/s12369-023-00968-4.

[16] M. Pinto-Bernal, M. Biondina, and T. Belpaeme, "Designing social robots with llms for engaging human interaction", *Applied Sciences*, vol. 15, no. 11, p. 6377, 2025. DOI: 10.3390/app15116377.

[17] A. Umbrico *et al.*, "A mind-inspired architecture for adaptive hri", *International Journal of Social Robotics*, vol. 15, no. 3,

pp. 371–391, 2023, ISSN: 1875-4805. DOI: 10.1007/s12369-022-00897-8.

[18] B. Ballinger, A. Engler, and M. A. Team, *Mycroft precise: A data-driven wake word engine*, GitHub repository, Open-source wake word detection engine for speech interfaces, 2018.

[19] S. Team, *Silero vad: Pre-trained enterprise-grade voice activity detector*, GitHub repository, Pre-trained voice activity detection model for real-time speech detection, 2024.

[20] M. Bain, J. Huh, T. Han, and A. Zisserman, "Whisperx: Time-accurate speech transcription of long-form audio", *INTERSPEECH 2023*, 2023.

[21] N. Reimers and I. Gurevych, *Sentence-bert: Sentence embeddings using siamese bert-networks*, 2019. arXiv: 1908.10084 [cs.CL].

Self-Competitive Simplification: Competition between Forward and Backward Simplification in Multi-Layered Neural Networks

Ryotaro Kamimura
Tokai University
Kitakaname, Hiratsuka, Kanagawa 259-1292, Japan
e-mail: ryotarokami@gmail.com

***Abstract*—The present paper aims to demonstrate that strong forces of simplification exist within neural networks. These forces compete with one another to make the simplification process as effective as possible. As a first approximation, we consider four types of simplification forces: forward, backward, collective, and individual simplification. The winners in the competition among these forces can efficiently simplify the network configuration, whereas the losers can eventually be utilized to introduce a certain level of complexity, which is required in actual learning. The proposed method was applied to a simple artificial data set containing both linear and non-linear inputs. When backward and forward simplification were forced to compete, forward simplification became more efficient in achieving simplification, while backward simplification played a complementary role by introducing additional complexity for improved generalization. These results suggest that, at a deeper level, multiple simplification forces coexist within neural networks and compete with one another to achieve efficient simplification.**

Keywords-competition; simplification; complexity; forward; backward; collective; individual; cost.

I. Introduction

This section explains the concept of self-competitive simplification, along with a brief introduction to the actual computational procedures.

A. Self-Learning

The present paper aims to demonstrate that neural networks have a strong intrinsic tendency toward simplification, which should ideally be observed even in the absence of external information. Previous studies on neural networks have primarily focused on representing input patterns as faithfully as possible, including efforts to obtain representations that lead to improved generalization. In contrast, the present study emphasizes the necessity of examining neural networks not only in relation to input patterns but also in terms of their internal structure and dynamics. This perspective suggests that we should explore the inner workings of neural networks, which are assumed to operate even without external information. This internal perspective is referred to as "self-learning," in which a network is ideally configured without external information, or at least self-organizes when triggered by only a very small amount of external input. When attention is shifted from external information to internal information, it becomes easier to observe how a network operates and, in particular, to identify the fundamental limitations of neural networks.

B. Self-Competitive Simplification for Complexity

At first glance, the simplification principle underlying self-learning appears to contradict the complexity required in network configurations during actual learning. To address this apparent contradiction, we hypothesize that such complexity arises from the coexistence of many different types of simplification procedures within a network. The simplification principle attempts to achieve simplification through all possible means, thereby producing various types of components and computational procedures that are suitable for simplification. These procedures compete with each other to make the network configuration as simple as possible. Some components or procedures win this competition, while others lose. Competition naturally produces losers with respect to simplification; however, from the perspective of actual learning, which requires complexity, these losing components may acquire external information and thereby increase complexity. Severe competition among multiple simplification procedures eventually creates room for the introduction of complexity. In this sense, losers in self-competition can become winners in acquiring external information during actual learning. It is important to note that, beneath the apparently complex network configurations observed at the surface level, there exists a simpler configuration at a deeper level, formed as a result of intense competition among diverse simplification procedures.

C. Self-Competitive Procedures

The competition among simplification procedures described above is referred to as "self-competition." This means that competition occurs not between input patterns but among components and learning procedures within the network itself. As a first approximation, we consider four competitive procedures: forward, backward, collective, and individual. In practical situations, such competition is assumed to be triggered by a small amount of external input information. First, competition is assumed to occur between forward and backward information processing. Information is transmitted from input to output in a forward manner, while it can also be propagated backward from output to input. These two modes of information flow compete with each other to facilitate both simplification and error minimization during learning.

In addition, components can be treated either collectively or individually. For example, a set of connection weights or neurons can be treated as a group, while each neuron or weight

 ISBN: 978-1-68558-375-0

can also be considered individually. Collective and individual procedures therefore compete with each other to achieve a simplified network configuration. Within this framework, all components and procedures compete, and those that win the competition are utilized more explicitly for simplification. Conversely, procedures that lose the competition for simplification may still be effective in reducing training error, even if they do not directly contribute to simplifying the network configuration.

D. Paper Structure

The rest of the paper is structured as follows. In Section Ⅱ, we discuss related work on competitive learning. In Section Ⅲ, we explain how to compute total and forward simplification, as well as the contradiction between collective and individual potentiality. In Section Ⅳ, we apply the method to a controlled data set containing both linear and nonlinear relationships. The results show that forward simplification wins the competition and produces efficient simplification. Backward simplification loses the competition but plays a role in introducing the complexity necessary for learning. The combination of forward and backward simplification can be used to achieve simplification while retaining sufficient complexity for effective learning.

II. Related Work

Related to conventional and well-established competitive learning, the method presented here introduces a new concept of competition, which ideally minimizes the influence of external information.

Competitive learning has been regarded as one of the fundamental learning paradigms since the early days of neural network research [1]–[3]. Numerous studies have aimed to extract representative features from input patterns [4]–[9]. Competitive learning has also been implemented as a core mechanism in Self-Organizing Maps (SOMs), a major unsupervised learning framework [10], [11].

In conventional competitive learning, a neuron that best represents the input patterns, often referred to as the best matching neuron, is selected. Such competition focuses on determining which neuron most closely matches the input, and thus which input features should be represented. In this sense, competitive learning, including SOMs, is guided by a simplification principle that aims to represent input patterns as economically as possible within a restricted network structure.

In contrast, we propose that competition should occur not only among input patterns but also among components and learning procedures within the network itself. This implies that competition can exist even without input patterns, at least ideally. We therefore refer to this framework as "self-competition," distinguishing it from conventional competitive learning. In self-competition, components compete with one another to achieve simplification from the very beginning of network formation, either without input patterns or with minimal external information.

At first glance, neural networks may appear to exhibit no explicit competition in their initial configurations. We hypothesize that what is observed is a collection of networks with many peripheral components and procedures, behind which there exists a simpler network in which all components and learning procedures compete to simplify the network configuration as much as possible. To reveal this internal structure, it is necessary to artificially simplify superficial and complex network configurations.

III. Theory and Computational Methods

In this section, after explaining the principles of simplification and competition, we introduce total and forward simplification.

A. Simplification Principle and Competition

The present paper introduces competition among components of neural networks, such as weights and neurons. In addition, computational procedures—including forward, backward, collective, and individual operations—are assumed to compete with each other in order to achieve maximum simplification.

We begin by explaining forward and backward simplification. Figure 1 illustrates the processes of forward and backward simplification. Figure 1(b) shows one of the final configurations obtained through forward simplification. In forward simplification, information entering a neuron is distributed, or de-compressed, to all neurons in the subsequent layer, and this process continues through all layers. Figure 1(d) illustrates the process of backward simplification, in which information in a layer is distributed to all neurons in the corresponding lower layer. We assume that competition between forward and backward simplification occurs, as shown in Figure 1(c).

Collective and individual competition are then applied either to a set of connection weights or to individual connection weights. In forward simplification, as shown in Figure 1(b), a neuron is regarded as being composed of a set of connection weights from all neurons in the lower layer. These neurons, represented as a set of connection weights, should be used as equally as possible, which is referred to as "collective" competition or simplification. In contrast, individual connection weights from neurons in the lower layer to a neuron in the subsequent layer compete with each other, which is referred to as "individual simplification." Backward simplification operates in the same manner as forward simplification.

B. Total Simplification

For simplicity, we consider only a single layer, because the same formulation can be applied to all other layers. The individual potentiality of a connection weight from the jth neuron to the kth neuron is computed as

$$u_{jk} = \mid w_{jk} \mid . \tag{1}$$

As mentioned above, all final values defined in this section can be obtained by averaging over all layers, including the input and output layers. In addition, for simplicity, it is assumed that the strength of all weights is greater than zero.

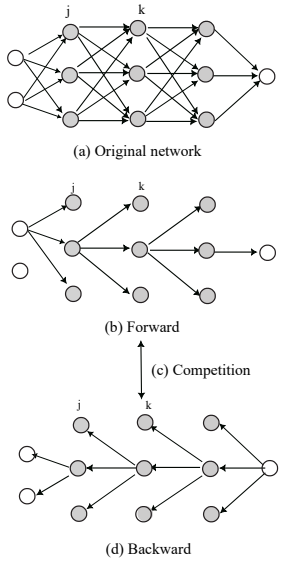


Figure 1. An original network (a), final configurations obtained by forward simplification (b) and backward simplification (d), and their competition (c).

The primary objective of simplification is to reduce the total cost potentiality of a network. The cost, or cost potentiality, is defined as the sum of individual potentialities:

$$C = \frac{1}{n_j n_k} \sum_{jk} u_{jk}, \tag{2}$$

where n_j and n_k denote the numbers of neurons in the corresponding layers.

Simplification aims to reduce this cost potentiality, which can be achieved in various ways.

C. Forward Simplification

We now formulate forward simplification. Collective simplification aims to simplify a group of connections from one layer to the subsequent layer. The potentiality of the kth neuron is defined as the sum of the absolute strengths of the incoming weights:

$$u_k = \sum_j u_{jk}. \tag{3}$$

The potentiality of the kth group is then computed as

$$h_k = \frac{u_k}{\max_{k'} u_{k'}}. \tag{4}$$

The collective potentiality is defined by

$$H = \frac{1}{n_k} \sum_k h_k. \tag{5}$$

Individual potentiality is defined for each connection weight. By normalizing the absolute weight by its maximum value, the relative individual potentiality is given by

$$g_{jk} = \frac{u_{jk}}{\max_{j'} u_{j'k}}. \tag{6}$$

Summing over all layers yields the individual potentiality:

$$G = \frac{1}{n_j n_k} \sum_{jk} \frac{u_{jk}}{\max_{j'} u_{j'k}}. \tag{7}$$

We define the contradiction between collective and individual potentialities as

$$D = H - G. \tag{8}$$

As this contradiction increases, collective potentiality increases while individual potentiality decreases, indicating a growing contradiction between connection weights and neurons. This quantity is not guaranteed to be positive; when individual potentiality increases and collective potentiality decreases, the contradiction can become negative.

This contradiction can be further extended by introducing the cost potentiality. The objective is to increase the contradiction D while minimizing the associated cost, leading to the contradiction ratio

$$R = \frac{D}{C}. \tag{9}$$

An increase in this ratio indicates a stronger contradiction between collective and individual potentiality accompanied by a reduction in total cost.

IV. Results and Discussion

After briefly explaining the overall experimental procedures, we present a numerical summary, followed by the results on collective and individual potentiality, cost and generalization, weights for all layers, and layer potentiality.

A. Experimental Outline

For clearly and simply demonstrating the performance of the simplification method, an artificial data set was used, in which both linear and non-linear relationships were implemented. The number of input patterns was 1000, and the number of inputs was seven. Among these inputs, the first four were linearly related to the targets, whereas the remaining three were non-linearly related to the targets. The non-linear inputs were generated by applying the square, sine, and logarithmic functions to the original inputs. The number of hidden layers was ten, and all parameter values, except for the number of learning steps (epochs), were set to the default values in the PyTorch package to facilitate easy reproduction of the results.

Figure 2 shows an outline of our experiments. As shown in Figure 2(a), initially, one step of collective simplification and four steps of individual simplification were applied to both forward and backward simplification. As a result, the final network was compressed into a network without hidden layers, as shown in Figure 2(b). This compressed network was compared with a prototype network in terms of correlation coefficients, which were computed using the data sets shown in Figure 2(d) and (e). When both backward and forward methods were used, backward simplification was first applied for 1100 learning steps, followed by forward simplification. This setting was chosen to improve generalization performance.

The main findings can be summarized as follows:

- The results show that forward contradiction maximization tended to win the competition over backward contradiction maximization.

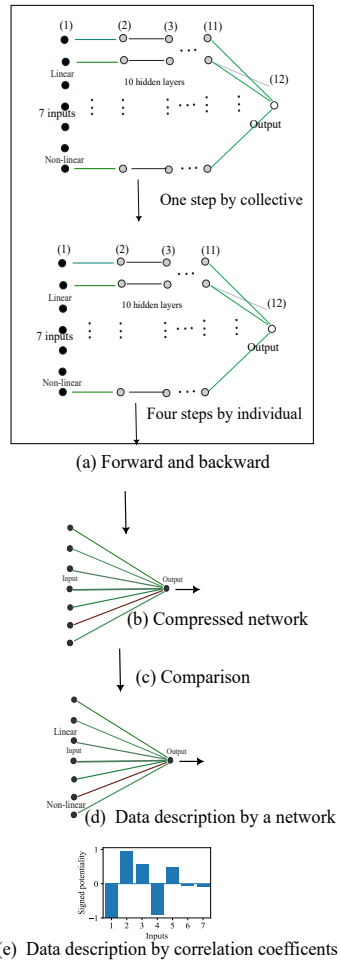


Figure 2. Forward and backward simplification (a), compressed network (b), comparison between compressed and prototype networks (c), data description by a network (d), and data description by correlation coefficients (e).

- Forward competition attempted to de-compress input information. Although input information tended to be compressed in the hidden layers, it could be de-compressed and fully transferred to all neurons in the subsequent layers.
- Backward competition focused on the output layer and target information. By losing the competition, backward simplification could be used to add an appropriate level of complexity, leading to improved generalization.
- Ultimately, the combination of forward and backward simplification could be used to simplify compressed networks by reducing the number of important inputs.

B. Numerical Summary

In the first place, we explain a summary of our experimental results, particularly from the viewpoint of generalization performance, since generalization has been one of the major indices of improved performance. However, it should be noted that this paper does not aim to explain the mechanism of improved generalization itself, but rather to clarify how different simplification procedures compete with each other during the simplification process.

The results show that forward simplification produced the most efficient simplification. In this case, the collective potentiality was the largest (0.737), the individual potentiality was the smallest (0.277), and correspondingly the largest difference, or contradiction, between collective and individual potentiality (0.460) was obtained. In addition, the ratio of this difference to the cost was the highest (2.931). This indicates that the forward type of simplification could most effectively simplify the network configuration in terms of both collective and individual simplification. When the two types of simplification forces were combined, the best generalization performance was obtained (0.861). In this combination, the difference or contradiction between collective and individual potentiality became negative (-0.078) in the backward simplification, indicating that the individual simplification force was stronger than the collective simplification force. At the same time, in this combination, the collective potentiality produced by the forward method was the third largest (0.688), the individual potentiality was the second smallest (0.279), and the resulting difference was the second largest (0.409). In addition, the ratio of this difference to the cost was the second largest (2.863). This suggests that, through the combination of forward and backward simplification, forward simplification was slightly weakened due to competition between the two simplification processes. This weakening could increase complexity and thereby improve generalization.

TABLE I. SUMMARY OF RESULTS, BASED ON MAXIMUM TESTING ACCURACY.

	Step	Coll	Indi	C-I	Cost	C-I/Cost	Accu
Conv	535	0.703	0.502	0.201	0.228	0.882	0.763
For	3316	**0.737**	**0.277**	**0.460**	0.157	**2.931**	0.839
Back	3241	0.547	0.363	0.184	0.174	1.057	0.851
FB(F)	**3632**	0.688	0.279	0.409	**0.143**	2.863	**0.861**
FB(B)		0.313	0.391	-0.078	**0.143**	-0.548	

C. Collective and Individual Potentiality

The results show that the forward potentiality tended to win the competition over the backward potentiality.

Figure 3 shows the collective (left), individual (middle), and the difference between them (right) as functions of the number of learning steps (epochs). When the conventional method shown in Figure 3(a) was used, the collective potentiality, the individual potentiality, and their difference or contradiction remained unchanged throughout the entire learning process. Naturally, the conventional method did not achieve simplification. Figure 3(b) shows the results obtained using the forward collective and individual method. The collective potentiality (left) remained higher throughout all learning steps, while the individual potentiality (middle) decreased clearly. As a result, the difference between them (right) increased and remained relatively high. Figure 3(c) shows the results obtained using the backward collective and individual method. The collective potentiality (left) was high at the beginning, but it eventually decreased. The individual potentiality (middle) decreased gradually. Consequently, the corresponding difference between them (right) increased slightly at the beginning and then tended to decrease slightly. This indicates that, with the backward method, the collective potentiality could not be increased,

which caused a reduction in the difference between collective and individual potentiality.

Figure 3(d) and (e) show the results obtained when the backward method was used during the first 1100 learning steps and the forward method was applied during the remaining learning steps. The final results were then computed in terms of potentialities using the forward method (d) and the backward method (e). Figure 3(d) shows the results obtained by the forward and backward method, where the potentialities were computed using the forward method. The collective potentiality remained higher even in the later stages of learning. The individual potentiality decreased fully, in the same manner as when only the forward method was used, as shown in Figure 3(b). The contradiction initially decreased due to the application of backward simplification, and then increased rapidly at the end as a result of forward simplification. In terms of backward potentialities shown in Figure 3(e), the collective potentiality (left) increased slightly during the backward computation period, namely up to 1100 learning steps. It then decreased considerably toward the end. The individual potentiality (middle) remained almost constant throughout the entire learning process. As a result, the difference between them increased at the beginning but decreased substantially at the end. This indicates that competition for simplification favored the forward method. In particular, the backward collective potentiality decreased considerably, implying that neurons were not used equally in terms of backward potentiality.

The combined use of forward and backward methods shows that the forward method won the competition for simplification. In this case, all neurons in the subsequent layers tended to be used equally, and the connection weights to those neurons were used locally. This implies that input information tended to be compressed within a layer, but could be de-compressed in the subsequent layer.

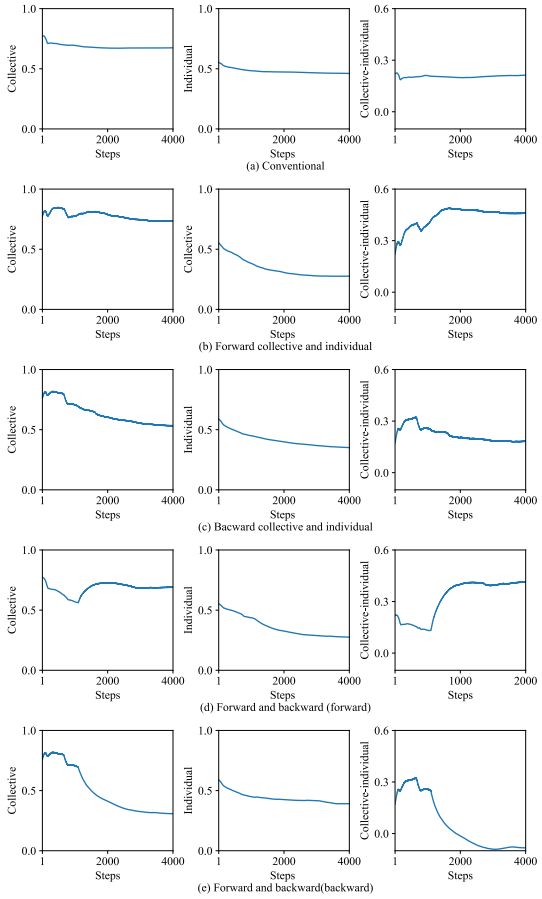


Figure 3. Collective (left), individual (middle) potentiality, and difference or contradiction between collective and individual (right) as a function of the number of learning steps (epochs) by the conventional method (a), by using the forward (b), by the backward (c) and by the forward and backward simplification with forward potentiality values (d) and backward potentiality values (e).

D. Cost and Generalization

Figure 4 shows the cost (left), the ratio of the difference between collective and individual potentiality to the cost (middle), and the generalization accuracy (right). Forward simplification considerably increased the contradiction ratio, whereas backward simplification could not increase it. When both methods were combined, this tendency was maintained in terms of forward simplification, but it was weakened in terms of backward simplification. This indicates that forward simplification won the competition over backward simplification.

Figure 4(a) shows the results obtained using the conventional method without potentiality control. The cost (left) increased rapidly, while the ratio remained constant. The generalization accuracy was relatively low and remained almost constant throughout the entire learning process. Figure 4(b) shows the results obtained using forward potentiality control. The cost (left) was kept small throughout all learning steps, and the ratio of the difference to the cost increased and approached 3. In addition, the generalization accuracy (right) increased gradually toward the end of learning. Figure 4(c) shows the results obtained using the backward method. The cost (left) was also relatively small, and the ratio (middle) initially increased but gradually decreased. The generalization accuracy (right) continued to increase until the end of learning and was slightly higher than that obtained by the forward method shown in Figure 4(b). Figure 4(d) shows the results obtained using the forward and backward method, where the potentialities were computed using the forward approach. The cost (left) was smaller than that of the conventional method, and the ratio (middle) increased rapidly and approached 3. The generalization accuracy (right) increased rapidly and reached the highest value, as summarized in Table I. Figure 4(e) shows the results obtained using the forward and backward method, where all potentialities were computed using the backward approach. One of the major findings is that the ratio of the difference to the cost (middle) decreased substantially. As explained in Figure 3(d), the backward collective potentiality was reduced. Nevertheless, the generalization performance increased gradually as the number of learning steps increased. The results demonstrate that the forward method could sufficiently increase the ratio of contradiction to the cost while improving generalization. In contrast, the backward method could not increase this ratio, even though the cost

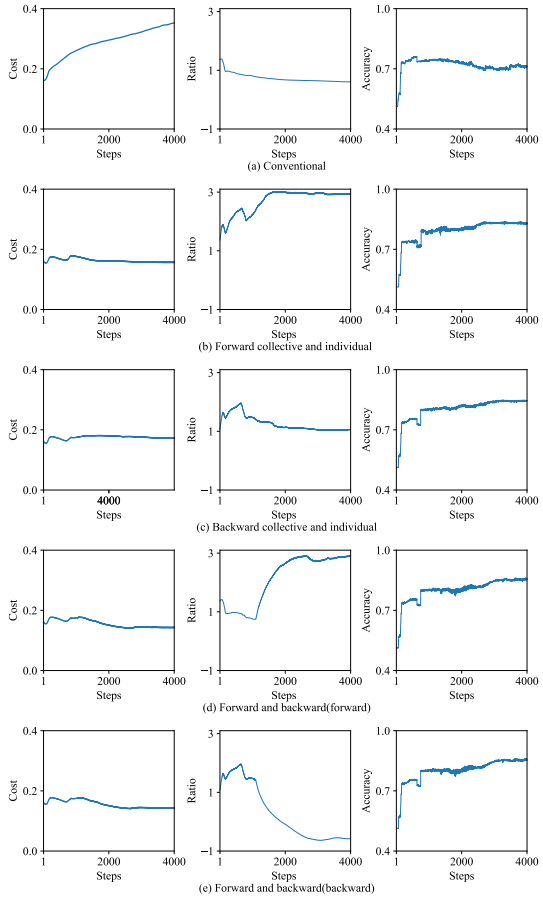


Figure 4. Cost (left), ratio of contradiction potentiality to cost (middle), and generalization accuracy (right) as a function of the number of learning steps by the conventional method (a), by using the forward collective and individual (b), by the backward collective and individual (c) and by the forward and backward with values, computed by the forward way (d) and the backward way (e).

was sufficiently small. This indicates that the forward method won the competition in simplification, whereas the backward method lost the competition, despite achieving relatively high generalization performance. This further suggests that simplification was mainly achieved by the forward method, while the backward method applied at the early stage of learning contributed to improved generalization.

E. Weights for All Layers

Figure 5 shows the connection weights after learning was completed using four different methods. Forward simplification produced more explicit patterns in the weights. When forward and backward simplification were combined, the resulting weights exhibited mixed properties, combining characteristics of both forward and backward simplification.

Figure 5(a) shows the results obtained using the conventional method. As can be seen in the figure, the weights appear to be randomly distributed, and only in the higher layers some regularities can be observed. Figure 5(b) shows the results obtained using the forward method. The weights were compressed in the majority of layers, whereas in the higher layers, compression and de-compression were mixed. Figure 5(c) shows the results obtained using the backward

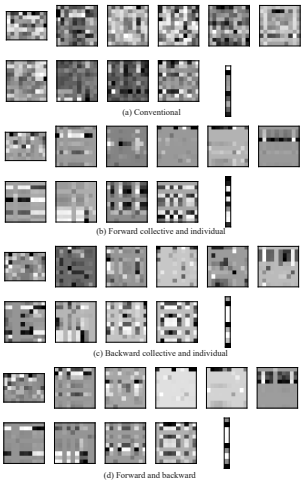


Figure 5. Connection weights across all layers obtained using the conventional method (a), the forward method only (b), the backward method only (c), and the combined forward and backward method (d).

method. The weights became sparser than those obtained using the conventional method, but they were more randomly distributed than those obtained using the forward method in Figure 5(b). Figure 5(d) shows the results obtained using the combined forward and backward method. As expected, mixed weight patterns were obtained, in which the characteristics of forward and backward simplification were combined. Overall, the weights became slightly more randomized than those obtained using the forward method alone.

By using both forward and backward simplification, the weights produced by forward simplification became slightly less explicit, introducing additional complexity into the connection weights. This added complexity may be one of the main factors contributing to improved generalization.

F. Layer Potentiality

Figure 6 shows the layer potentiality, computed by summing all individual potentialities within each layer. Forward simplification exhibited a clearer pattern, in which the layer potentialities decreased initially and then increased toward the end. When both forward and backward simplification were applied, this tendency was slightly attenuated.

Figure 6(a) shows the results obtained using the conventional method. The layer potentiality initially exhibited a nearly uniform distribution. Gradually, the layer potentiality of the output layer became the largest, indicating that the output layer played the most important role in learning. Figure 6(b) shows the results obtained using the forward method with collective and individual simplification. The layer potentiality was relatively large at the beginning, then became smallest in the middle layers, and finally became largest at the output layer. Figure 6(c) shows the results obtained using the backward method. The layer potentialities were almost the same as those obtained using the forward method in Figure 6(b), but their overall magnitudes were weaker. Figure 6(d) shows the results obtained when the forward and backward methods were combined. The strength of the layer potentialities reflected

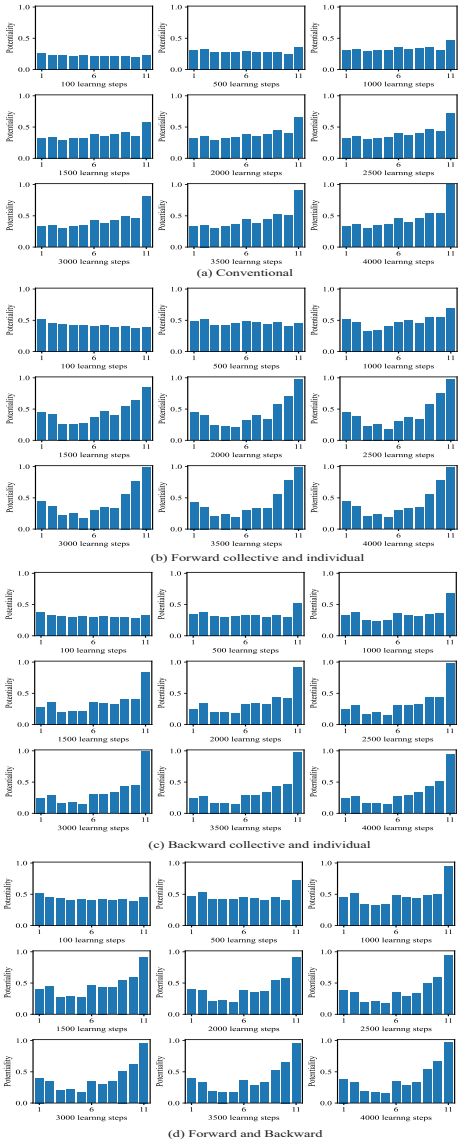


Figure 6. Layer potentialities across all layers, defined as the sum of individual potentialities within each layer, obtained using the conventional method (a), the forward collective and individual method (b), the backward collective and individual method (c), and the combined forward and backward method (d).

a mixture of the characteristics observed in the forward and backward methods.

When both forward and backward simplification were used, the layer potentialities became slightly weaker than those obtained using only the forward method. This is one of the main reasons for the improved generalization, because additional complexity was introduced through the effect of backward simplification.

V. Conclusion and Future Work

The present paper aimed to demonstrate the existence of a simplification principle in which multiple simplification procedures compete with each other. In this study, forward and backward simplification competed with one another, and both collective and individual simplification were incorporated into the learning process. The results show that forward simplification, which aims to simplify the network configuration from the viewpoint of input information, won the competition. As a result, input information was compressed and subsequently de-compressed across the layers. The backward simplification, which lost the competition, could instead be utilized to introduce an appropriate level of complexity, leading to improved generalization. Competition among components and learning procedures produces winning mechanisms through which simplification can be further deepened. Conversely, the losing mechanisms can play a complementary role by adding complexity, for example, to enhance generalization performance.

The present paper is a preliminary study on the competition between forward and backward simplification. Further investigation is required to examine the competitive effects of forward and backward simplification by more carefully controlling the associated simplification parameters. In addition, larger and more practical data sets should be employed to more accurately evaluate the performance of the proposed method.

References

[1] D. E. Rumelhart and D. Zipser, “Feature discovery by competitive learning,” *Cognitive Science*, vol. 9, pp. 75–112, 1985.

[2] C. S. Choy and W. Siu, “A class of competitive learning models which avoids neuron underutilization problem,” *IEEE Transactions on Neural Networks*, vol. 9, no. 6, pp. 1258–1269, 1998.

[3] A. Banerjee and J. Ghosh, “Frequency-sensitive competitive learning for scalable balanced clustering on high-dimensional hyperspheres,” *Neural Networks, IEEE Transactions on*, vol. 15, no. 3, pp. 702–719, 2004.

[4] D. Krotov and J. J. Hopfield, “Unsupervised learning by competing hidden units,” *Proceedings of the National Academy of Sciences*, vol. 116, no. 16, pp. 7723–7731, 2019.

[5] Y. Lu, Y.-M. Cheung, and Y. Y. Tang, “Self-adaptive multiprototype-based competitive learning approach: A k-means-type algorithm for imbalanced data clustering,” *IEEE Transactions on Cybernetics*, vol. 51, no. 3, pp. 1598–1612, 2019.

[6] T. Shinozaki, “Biologically motivated learning method for deep neural networks using hierarchical competitive learning,” *Neural Networks*, vol. 144, pp. 271–278, 2021.

[7] M. G. Mahdy, A. R. Abas, and T. M. Mahmoud, “Multi-phase adaptive competitive learning neural network for clustering big datasets,” in *The International Conference on Artificial Intelligence and Computer Vision*, Springer, 2021, pp. 731–741.

[8] G. Lagani, F. Falchi, C. Gennaro, and G. Amato, “Training convolutional neural networks with competitive hebbian learning approaches,” in *International Conference on Machine Learning, Optimization, and Data Science*, Springer, 2021, pp. 25–40.

[9] P. Li, S. Tu, and L. Xu, “Deep rival penalized competitive learning for low-resolution face recognition,” *Neural Networks*, vol. 148, pp. 183–193, 2022.

[10] T. Kohonen, “The self-organizing maps,” *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990.

[11] Y. Xu, L. Xu, and T. W. Chow, “Pposom: A new variant of polsom by using probabilistic assignment for multidimensional data visualization,” *Neurocomputing*, vol. 74, no. 11, pp. 2018–2027, 2011.

Cognitive Foundations of Real-Time Language Communication: Toward a Theoretical Framework of Behavioral Linguistics

Muneo Kitajima
Nagaoka University of Technology
Nagaoka, Niigata, Japan
Email: mkitajima@kjs.nagaokaut.ac.jp

Makoto Toyota
T-Method
Chiba, Japan
Email: pubmtoyota@mac.com

Jérôme Dinet
Université de Lorraine
Nancy, France
Email: jerome.dinet@univ-lorraine.fr

Katsuko T. Nakahira
Nagaoka University of Technology
Nagaoka, Niigata, Japan
Email: katsuko@vos.nagaokaut.ac.jp

***Abstract*— This paper explores the cognitive processes involved in real-time language communication during conversations. It emphasizes the dynamic interplay between speakers and listeners, highlighting how both verbal and nonverbal cues contribute to effective communication. The purpose of this paper is to construct a theory of behavioral linguistics on the cognitive architecture, Model Human Processor with Realtime Constraints (MHP/RT), that explains how language is generated in real-time, addressing the limitations of traditional models based solely on Skinner's behavioral psychology for verbal development with System 1 or Chomsky's linguistic theory grounded in the development of formal grammar through System 2, and regarding current natural language being formed through the interaction of Chomsky's grammatical system built upon Skinner's foundation. The proposed theory of behavioral linguistics offers a comprehensive framework for understanding real-time language generation, integrating insights from cognitive psychology and behavioral economics. By framing language use as an action controlled by bounded rationality, this paper highlights the cognitive constraints that influence how individuals communicate. This perspective encourages a more nuanced understanding of conversational competence, suggesting that effective communication relies not only on linguistic knowledge but also on the ability to anticipate and adapt to the dynamic nature of interactions. The implications for improving communication in various contexts, such as education and therapy, underscore the relevance of this research in enhancing everyday interactions and reducing misunderstandings.**

Keywords- Behavioral Linguistics; Everyday Conversation; Verbal Behavior; Nonverbal Communication; MHP/RT, GOMS.

I. INTRODUCTION

For humans, social creatures who use language, everyday conversations with close friends and family bring richness to daily life and enable us to spend fulfilling and happy times. In conversation, the roles of speaker and listener alternate appropriately among participants, maintaining the flow of dialogue. Listeners prepare their own responses while listening to the speaker, anticipating when it will be their turn to speak next in the conversation. The speaker attempts to convey the information they wish to communicate to the listener through both linguistic information transmitted as auditory information and nonverbal information conveyed through visual information such as gestures and eye contact. Skinner [1] conducted a functional analysis of this verbal behavior. Behavior analysts have been working on developing ideas based on verbal behavior for fifty years, and despite this, experience difficulty explaining generative verbal behavior [2]. This suggests that explanations based solely on accumulated conversational information have limitations, and further implies that complex information processing may be involved.

In everyday conversation, speakers adjust their speech to suit the listener's state and deliver appropriate utterances at the right time. Speakers cannot fully grasp the listener's state of understanding, and it is thought that the content of speech is unconsciously selected from several candidates. In verbal behavior, the concepts of bounded rationality and the satisficing principle proposed by Simon [3] are at work. Kahneman proposed the dual-process theory as the cognitive basis for how bounded rationality operates during decision-making [4], laying the foundation for behavioral economics.

This paper proposes that the theory of behavioral linguistics, which can address the verbal behavior of real-time language generation in everyday conversation, can be constructed by examining it on the cognitive foundation of Model Human Processor with Realtime Constraints (MHP/RT) [5]–[7], which concerns action selection under real-time constraints based on an understanding that it is grounded in dual-process theory [4][8]. In MHP/RT, we consider that linguistic behavior emerges through two processes: System-2-Before-Event-Mode, where the listener consciously determines what to say in preparation for becoming the next speaker, and System-1-Before-Event-Mode, where the speaker unconsciously adjusts the content and manner of speech to fit the situation immediately before speaking. Much of verbal behavior is executed unconsciously in a "Goals, Operators, Methods, and Selection rules (GOMS)"-like manner. By employing a rich array of methods concerning mode transitions in the dual-process, it is thought possible to respond instantly to changes in the other party [9]. Furthermore, as a mechanism preventing conversational breakdown, we assume appropriate synchronization—weak synchronization [10]—for these modes

 ISBN: 978-1-68558-375-0

between the speaker and listener. Weak synchronization has been demonstrated as a mechanism for user immersion within virtual environments. In smooth everyday conversation, a state analogous to immersion is thought to manifest. This paper explains these mechanisms that form the foundation of behavioral linguistics.

The main objective of the paper is to propose a theory of behavioral linguistics to explain the real-time generation of language in everyday conversation. This paper is structured as follows. Section II explains the positioning of this research while citing related studies. Section III describes the Perceptual, Cognitive, and Motor (PCM) processes performed by listeners and speakers, and how memory is utilized. Section IV focuses on speaker turn-taking in everyday conversation, elucidating the mechanisms underlying smooth conversational flow. Section V summarizes this research and highlights its relevance to our daily lives.

II. Related Works

The modern scientific analysis of verbal behavior originates in the work of B. F. Skinner [1], whose verbal behavior proposed a functional taxonomy of language based on operant conditioning principles. Skinner's account emphasized environmental contingencies and reinforcement histories as determinants of linguistic behavior. While this framework proved influential, it has been criticized for its difficulties in explaining generativity and the flexible production of novel utterances in spontaneous conversation [11].

To address generativity within a behavior-analytic tradition, Steven C. Hayes and colleagues developed Relational Frame Theory (RFT) [12], which conceptualizes language as generalized relational responding. RFT offers a more powerful explanation of symbolic and rule-governed behavior than classical operant approaches. However, its primary focus lies in derived relational responding and symbolic transformation rather than in the real-time temporal coordination observed in everyday conversational turn-taking.

In cognitive psychology, bounded rationality theory introduced by Herbert A. Simon reframed human decision-making as adaptive under cognitive and environmental constraints [13][14]. Rather than optimizing, individuals select options that are "good enough" given limited time and information. This perspective is particularly relevant to conversational contexts, where speakers must rapidly select utterances without complete knowledge of the listener's internal state.

Building upon bounded rationality, Daniel Kahneman synthesized decades of research on dual-process theory, distinguishing between fast, automatic (System 1) and slow, deliberative (System 2) cognitive processes [15]. Earlier formulations of dual-process models (e.g., [8][16]) emphasized this distinction. While these frameworks have been influential in behavioral economics and decision science, they have not been systematically integrated into a theory of verbal behavior under real-time interactive constraints.

Within psycholinguistics, research on turn-taking demonstrates that listeners begin planning their responses before the current speaker finishes [17]. Conversation analysis has further shown that turn transitions are typically characterized by minimal gaps and overlaps [18]. These findings suggest highly efficient predictive and timing mechanisms, yet they are often treated independently of behavior-analytic theory or broader models of action selection.

In parallel, human performance modeling provides additional relevant foundations. The Model Human Processor (MHP) proposed by Card, Moran, and Newell [19] conceptualizes human activity in terms of Perceptual, Cognitive, and Motor (PCM) subsystems. The Goals, Operators, Methods, and Selection rules (GOMS) framework further formalizes procedural task execution. While these models offer powerful tools for analyzing structured task behavior, they have rarely been extended to spontaneous conversational language generation in naturalistic settings.

So, by integrating all these prior foundations, our paper is innovative for several reasons:

(a) Unlike classical behaviorism [1], the proposed framework incorporates internal action-selection mechanisms grounded in bounded rationality [13] and dual-process cognition [4], enabling an explanation of generative and adaptive utterance formation;
(b) While dual-process theory [15] provides a general cognitive distinction, our paper specifies its implementation within conversational micro-dynamics through two system modes;
(c) By situating verbal behavior within PCM loops [19], our model accounts for the integration of linguistic, nonverbal, and timing behaviors under real-time constraints;
(d) Drawing from findings in conversation analysis [18] and predictive planning research [17], the introduction of weak synchronization [10] provides a mechanistic explanation for smooth turn-taking and breakdown prevention;
(e) Finally, extending traditional MHP and GOMS approaches [19], our proposed MHP/RT [5]–[7] framework explicitly models linguistic action selection under temporal pressure, bridging human performance modeling and behavioral linguistics.

III. Cognitive Process Model of Conversation

A. *Description and Scope of the Targeted Conversation Situation*

This study focuses on smooth conversation (dialogue) between two individuals. In this conversation, both verbal and nonverbal communication occur. In the former, information is transmitted between participants through auditory information conveyed by sound waves. In the latter, information is transmitted through information from each modality conveyed via the five senses (vision, hearing (including filler information such as interjections), touch, smell, and taste).

This study focuses on verbal communication to delve deeply into the conversation itself. In verbal communication, the observable states of interlocutors can be divided into "speaking" and "not speaking." Therefore, the possible states are as follows:

1) One party is speaking, while the other is not;

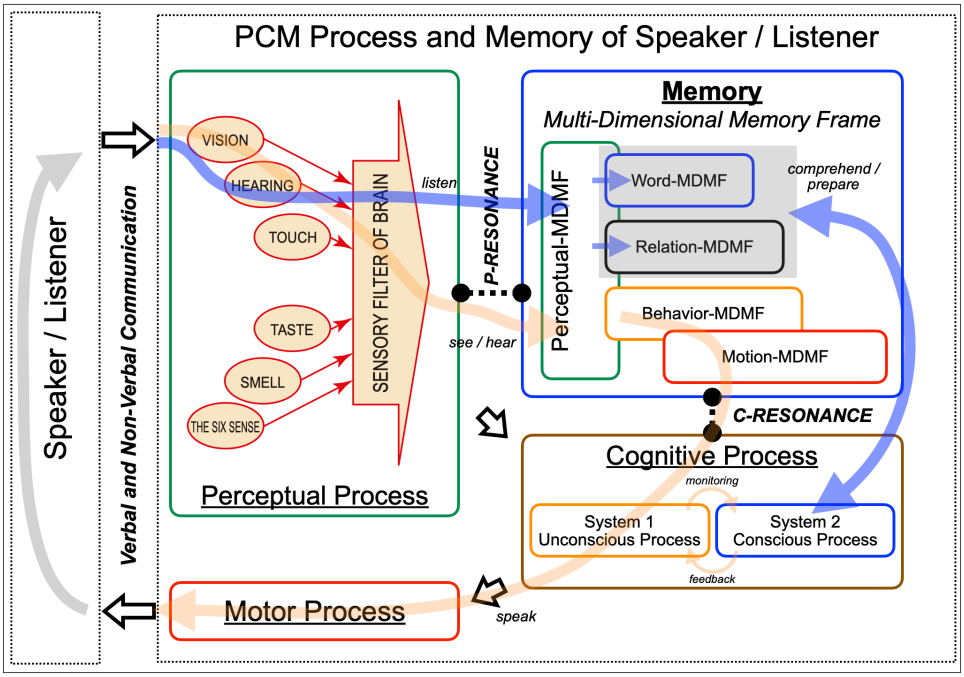


Figure 1. PCM processes and the Multi-Dimensional Memory Frame of conversation participants (modified from [20, Figure 1]).

2) Neither party is speaking;
3) Both parties are speaking simultaneously.

This study examines smooth conversation. In such conversations, both parties concentrate on the conversation and aim to achieve its purpose. Each should be performing verbal behaviors efficiently, without wasting time on unnecessary thought. Therefore, this study assumes the first situation described above. This means that conversation participants have the roles of "speaker" and "listener," and the conversation proceeds through alternating speakers. In this study, the timing of speaker change—that is, the moment when the event of speaker alternation occurs—is considered to be the instant when the speaker changes to the listener.

B. PCM Process Executed by Conversation Participants

We have published cognitive science analyses targeting seamless interaction with the environment [9][21][22]. In these analyses, the environment included Virtual Reality (VR) and other humans. These analyses were grounded in the MHP/RT, a cognitive architecture capable of simulating real-time action selection processes in daily life. This study views everyday conversation as a form of interaction between the environment and humans. By leveraging existing knowledge, we aim to deepen our understanding of the mechanisms enabling seamless conversation.

1) PCM Process and Memory: Conversation participants take on the roles of speaker or listener as the conversation progresses. The speaker actively engages in the conversation, while the listener passively participates. Speaker changes occur at appropriate times, with each person attempting to convey what they wish to communicate to others through verbal and nonverbal communication. When speakers and listeners interact through the format of conversation, sensory nerves within their sensory organs respond to the physical and chemical stimuli emitted by both parties, thereby incorporating information into the individual's brain. Each participant's brain acquires information about its own current activity through multiple sensory organs and generates bodily movements appropriate to the present situation.

Figure 1 illustrates MHP/RT and the process by which information emitted by a conversation participant is incorporated into the body via sensory nerves, undergoes information processing within the brain, and is then expressed as speech via motor nerves, thereby advancing the conversation. This process involves memory, modeled as Multi-Dimensional Memory Frame, and PCM processes. The cognitive process is a dual-process comprising unconscious and conscious processes. This is also referred to as Two Minds, where System 1 executes unconscious processes and System 2 executes conscious processes [4][15]. The Multi-Dimensional Memory Frame consists of Perceptual-, Behavior-, Motor-, Relation-, and Word-Multi-Dimensional Memory Frame. The Perceptual-Multi-Dimensional Memory Frame overlaps with the Behavior-, Relation-, and Word-Multi-Dimensional Memory Frame. This allows activity to propagate from the Perceptual- to Motor-Multi-Dimensional Memory Frame.

Perceptual information taken in from the environment through sensory organs *resonates* with information in the Multi-Dimensional Memory Frame, which is called P-Resonance [20]. In Figure 1, this process is indicated by the symbol •—•. Resonance occurs first in the Perceptual-Multi-Dimensional Memory Frame and activates the memory network. After that,

the activation spreads to the memory networks that overlap with the Perceptual-Multi-Dimensional Memory Frame, and finally to the Motor-Multi-Dimensional Memory Frame.

In cognitive processing based on the Two Minds framework [4][15], conscious processing (System 2) and unconscious processing (System 1) operate in an interrelated manner [20][23]. System 2 utilizes the Word- and Relation-Multi-Dimensional Memory Frame via C-Resonance, while System 1 draws on the Behavior- and Motor-Multi-Dimensional Memory Frame via the same mechanism. Motor sequences are then expressed according to the Motor-Multi-Dimensional Memory Frame. The memories involved in the production of actions are updated to reflect the traces of their use process and influence the future action selection process.

2) PCM Processing and Memory during Conversation: Using Figure 1, we describe processes occurring during conversation: the PCM process and propagation of activation within memory. Regardless of whether a conversation participant is acting as a speaker or listener, information taken into the brain via sensory organs activates the Perceptual-Multi-Dimensional Memory Frame through P-resonance.

The activation propagates within the Multi-Dimensional Memory Frame, undergoing processing via the following two pathways. The first pathway is indicated by the blue arrow in the figure. In this pathway, conscious information processing by System 2 is executed via C-resonance between information propagated from the Perceptual-Multi-Dimensional Memory Frame to Word- and Relation-Multi-Dimensional Memory Frame. The second pathway is indicated by the yellow arrow in the figure. Along this pathway, unconscious information processing by System 1 occurs via C-resonance with information propagated along the Perceptual-, Behavior-, and Motor-Multi-Dimensional Memory Frame. This latter process connects to motor processes, resulting in physical actions reflecting the active state of Motor-Multi-Dimensional Memory Frame.

Speakers and listeners execute the PCM process shown in Figure 1 using the Multi-Dimensional Memory Frame during conversation to perform the following tasks.

The tasks to be performed by the speaker are as follows:

- Utterance: Continuously speaks words based on the selected content;
- Adjustment: Adjusting one's speech content while observing the listener's reaction to one's own utterances or while listening to one's own utterances.

The tasks to be performed by the listener are as follows:

- Understanding: Comprehending the content of the other person's utterances. Here, both verbal and nonverbal information is utilized;
- Nonverbal responses: While listening to the other person's utterances and observing accompanying actions (nonverbal information), one exhibits unconscious nonverbal reactions (eye contact, facial expressions, gestures (nodding), interjections, etc.);
- Preparation: While listening to the other person's utterances, deliberate on what to say after the speaker alternation.

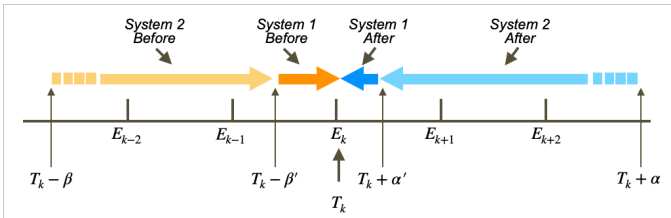


Figure 2. Four processing modes of MHP/RT.

C. Characterization of Conversational Behavior based on the Four Processing Modes

As shown in Figure 1, the PCM process is a cycle. When a speaker hears their own utterance, notices a mispronunciation, and corrects it, the process connects as follows: Motor (utterance) → Perception (listening) → Cognition (detecting the mispronunciation, deciding on a correction method) → Motor (utterance) → $\cdots$. In this way, the PCM process runs as a continuous cycle, but by breaking it at consciously recognized events, it can be perceived as a sequence of events. When considering a conversational behavior, a representative event is speaker alternation. Alternatively, one may become conscious of their own previous utterances while speaking, as if noticing a slip of the tongue. This too is an event occurring during conversation. By focusing on such events, the unbroken PCM cycle can be captured through the following four processing modes.

1) Four Processing Modes of MHP/RT : The experience associated with an individual's activity is characterized by a series of events that are consciously recognized serially. Let $E(T_k)$ denote the event that occurred at time T_k. The experience is then defined as a series of events along the timeline as follows:

$$\cdots \rightarrow E(T_{k-1}) \rightarrow E(T_k) \rightarrow E(T_{k+1}) \rightarrow \cdots$$

Considering the way System 1 and System 2 are involved in individual events, four processing modes can be defined as shown in Figure 2.

Let us focus on an event that occurs at time T_k. For an "event $E(T_k)$" that should occur at time T_k, there exist System 2 conscious processes and System 1 unconscious processes related to $E(T_k)$ before that time T_k. Also, for the "executed event $E(T_k)$" at time T_k, there exist unconscious processes of System 1 and conscious processes of System 2 involving $E(T_k)$ after that time T_k. MHP/RT's System 1 and System 2 operate before and after the event $E(T_k)$ in one of four processing modes for this event.

a) Before the Event ($T < T_k$) : The event $E(T_k)$ that occurs at time T_k reflects the result of the resonance (P-Resonance) between the Multi-Dimensional Memory Frame and the perceptual and cognitive systems—System 1 and System 2—during the time before T_k. $E(T_k)$ is generated by the activities of System 1 and System 2 in the time period before T_k. The different time bands of processing activities result in two processing modes before the event (corresponding

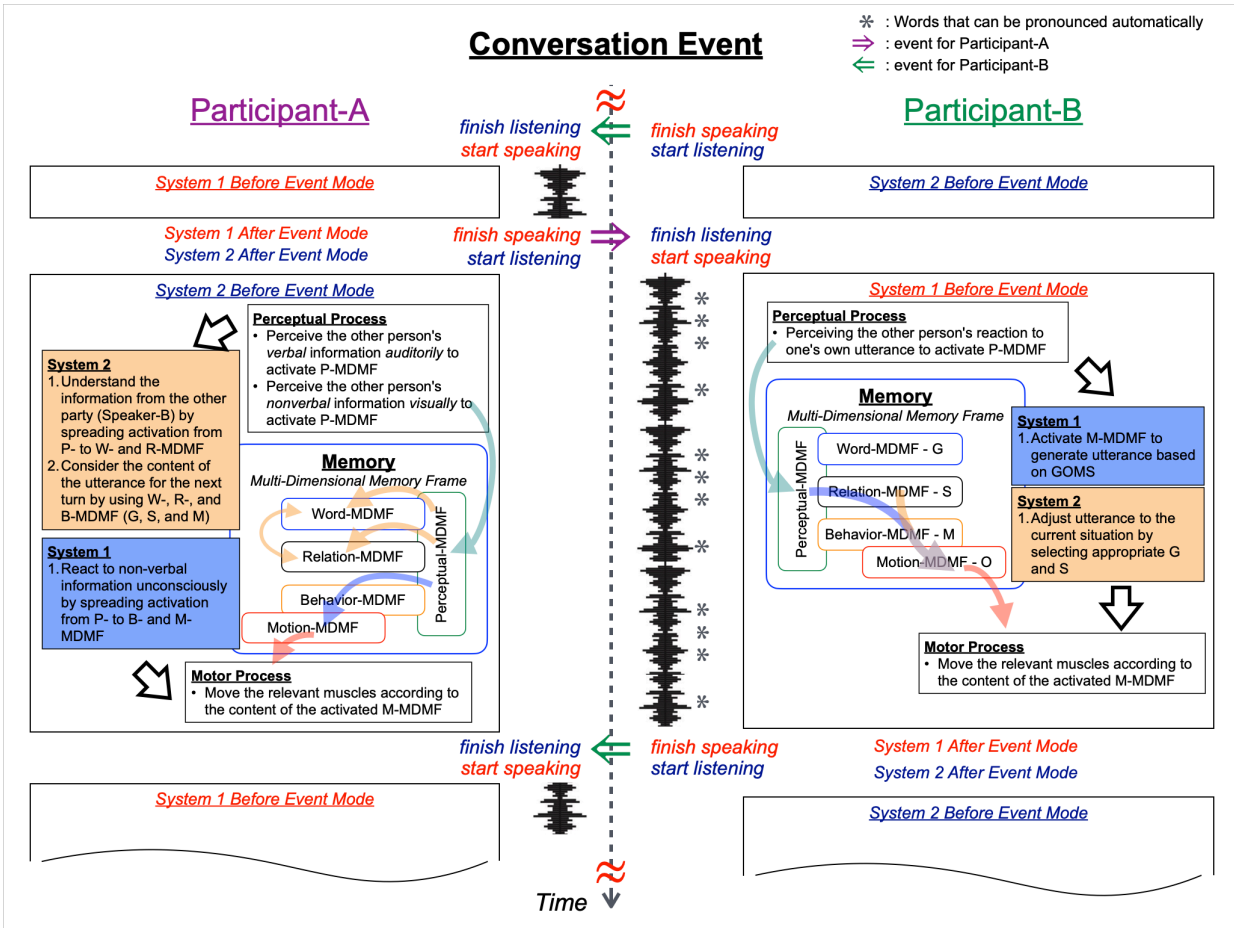


Figure 3. Basic PCM processes for listeners and speakers in conversation and the use of memory.

to the part before time T_k in Figure 2). The two processing modes are:

▷ [System-2-Before-Event-Mode]:
In the time range of $T_k - \beta \leq t < T_k - \beta'$, MHP/RT plans for future events to occur. There is enough time to think carefully.

▷ [System-1-Before-Event-Mode]:
In the time range of $T_k - \beta' \leq t < T_k$, the action selections smoothly generate the immediate event. Here, a series of action selections is executed through feedforward processing led by System 1. During this time, System 2 evaluates the results of the action selections in a timely manner. If it determines that the system is likely to deviate from the expected trajectory or has already deviated, it issues instructions to System 1 for trajectory correction.

Here, $\beta > \beta'$, $150\text{msec} < \beta' < T_k - T_{k-1}$, and β ranges from seconds to hours and months.

b) After the Event ($T > T_k$) : When event $E(T_k)$ occurs at time T_k, the result is stored. Actions occur by integrating the resonances that emerge through interacting with the environment prior to the event, and after the actions are taken, they are bundled and collected. The existing Multi-Dimensional Memory Frame are updated to reflect the results of $E(T_k)$ by the activities of System 1 and System 2 during the time period after T_k. The different time bands of processing activities result in two processing modes after the event (corresponding to the part after time T_k in Figure 2).

▷ [System-1-After-Event-Mode]:
In the time range of $T_k < t \leq T + \alpha'$, to perform better for the same event that may be encountered in the future, the connection between the incoming perceptual information and the output motor content is adjusted unconsciously.

▷ [System-2-After-Event-Mode]:
In the time range of $T_k + \alpha' < t \leq T_k + \alpha$, the event is reviewed and reflected upon. The results are stored and used in the next System-2-Before-Event-Mode before a similar event occurs.

The minimum value of α' is $\sim$ 150msec, and α ranges from seconds to months. In these two modes, action selection results for the event at T_k would be reflected in the network connections of the respective Multi-Dimensional Memory Frame.

2) Basic PCM Process for Conversational Behavior: Figure 3 shows the PCM process and memory during a conversation between conversation participant A and conversation participant B. The speaker executes utterance and adjustment,

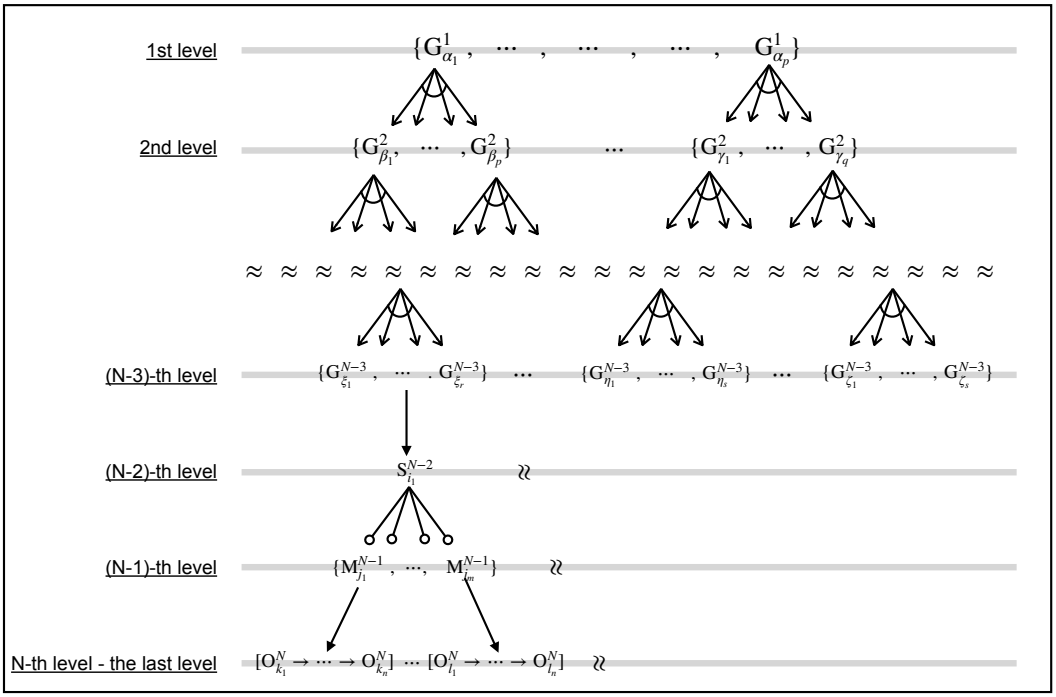


Figure 4. GOMS connection structure [23, Figure 8].

while the listener executes understanding, nonverbal response, and preparation by operating the PCM process (see Section III-B2).

In conversation, cognitive processes are executed through the four processing modes described in Section III-C1. The four processing modes are defined using the time at which an "event" occurs. Since conversations are organized and progress through timely speaker alternations, this study defines the basic event in conversation as "the time when the speaker finishes their utterance," as shown in Figure 3.

The left box in Figure 3 shows the process when conversation participant A is listening to conversation participant B's utterance. The right box in Figure 3 shows the process when conversation participant B is speaking. The processes described within these boxes are explained below.

a) Understanding the Content of Utterance: This process is described in System 2-1 in the left box of Figure 3. Propagation of activation within the Multi-Dimensional Memory Frame is indicated by green and yellow arrows. The speaker's utterance (verbal information) is incorporated as auditory information to activate the Perceptual-Multi-Dimensional Memory Frame. This activation propagates to the Word- and Relation-Multi-Dimensional Memory Frame, enabling understanding of the utterance content in System-2-Before-Event-Mode. The understanding is represented as an activation pattern within the Multi-Dimensional Memory Frame.

b) Preparation for Utterance: This process is described in System 2-2 in the left box of Figure 3. Propagation of activation within the Multi-Dimensional Memory Frame occurs between the Word- and Relation-Multi-Dimensional Memory Frame. This is indicated by the yellow arrows connecting them. Preparation of utterances is performed by System-2-Before-Event-Mode, utilizing the knowledge described below.

Smooth conversation is executed as a routine goal-oriented task. Therefore, this study represents the knowledge utilized during speech using the GOMS model, which consists of goals, operators, methods, and selection rules. Figure 4 illustrates the connection structure of G, O, M, and S. Layers 1 through $(N-3)$ correspond to the goal structure, layer $(N-2)$ to selection rules, layer $(N-1)$ to methods, and layer N to operators [19]. In the GOMS model, the goal structure G is stored in the Word-Multi-Dimensional Memory Frame, and the selection rules S, which determine the appropriate method to apply based on the situation, are stored in the Relation-Multi-Dimensional Memory Frame. These are objects consciously manipulated by System 2. The methods M, which are pointers to operator sequences, are stored in the Behavior-Multi-Dimensional Memory Frame. These are linked to the Relation-Multi-Dimensional Memory Frame. The operators O are stored in the Motor-Multi-Dimensional Memory Frame, and the operator sequences defined by the methods are passed to the motor process, where actions are executed [23].

Applying the general GOMS model to conversational behavior yields the following: the Word-Multi-Dimensional Memory Frame represents the information (goal) the speaker wishes to convey to the listener in utterances following turn-taking. The Behavior-Multi-Dimensional Memory Frame represents the utterance content (method), expressed as the sequence of words stored in the Motor-Multi-Dimensional Memory Frame that is generated automatically and unconsciously. The Relation-Multi-Dimensional Memory Frame represents the candidate selection rules (selection rules) for timely substitution of utterance content based on the listener's situation and the speaker's own situation [19][23].

In the GOMS model used for conversational behavior, the components related to utterance preparation (see Section III-C2b) are as follows. The activation patterns of the Multi-Dimensional Memory Frame generated by utterance understanding (see Section III-C2a) and activated in relation to nonverbal reactions are reflected in the conversation goal structure G within the activated Word-Multi-Dimensional Memory Frame and the selection rules S within the Relation-Multi-Dimensional Memory Frame as the conversation progresses. This process effectively updates the activation state of the GOMS connection structure and places the utterance method candidates stored in the Behavior-Multi-Dimensional Memory Frame to be executed after speaker turn-taking into a standby state.

c) Nonverbal Reaction: This process is described in System 1-1 in the left box of Figure 3. The propagation of activation within the Multi-Dimensional Memory Frame is indicated by green, blue, and red arrows. The listener perceives nonverbal information—facial expressions, gaze, blinking, nodding, touch, murmurs, etc.—emitted by the speaker during utterance through the five senses as perceptual information, thereby activating the Perceptual-Multi-Dimensional Memory Frame. Simultaneously, the Perceptual-Multi-Dimensional Memory Frame overlaps with the Word- and Relation-Multi-Dimensional Memory Frame, which are activated by speech understanding and speech preparation occurring in System-2-Before-Event-Mode. Therefore, Perceptual-Multi-Dimensional Memory Frame activation occurs through these two pathways. The listener exhibits nonverbal behavior via System 1 that reflects the propagation of activation from the Perceptual- to Behavior- and Motor-Multi-Dimensional Memory Frame.

d) Generation of Utterance: This process is described in System 1-1 in the right box of Figure 3. The propagation of activation within the Multi-Dimensional Memory Frame is indicated by the yellow arrows. One candidate suitable for the current situation is selected by applying the selection rules to the speech method candidates in the Behavior-Multi-Dimensional Memory Frame that are placed in a standby state during speech preparation while listening. Then activation is propagated to the Motor-Multi-Dimensional Memory Frame. Following the activation pattern, the motor process is activated to produce speech (red arrow). This process is executed by System-1-Before-Event-Mode. Upon utterance completion, the utterance becomes conscious as an event, and then System-1-After-Event-Mode and System-2-After-Event-Mode are executed. In System-2-After-Event-Mode, the completed goal is deactivated, and processing moves to the next stage: System-2-Before-Event-Mode as listener.

e) Adjustment of the Contents of Utterance: This process is described in System 2-1 in the right box of Figure 3. The propagation of activation within the Multi-Dimensional Memory Frame is indicated by green and blue arrows. The speaker's own utterance is input auditorily and activates the Perceptual-Multi-Dimensional Memory Frame. Furthermore, the listener's nonverbal responses are also perceived through the five senses, activating the Perceptual-Multi-Dimensional Memory Frame. From there, activation propagates to the Word-, Relation-, and Behavior-Multi-Dimensional Memory Frame.

One's own utterances are generated reflecting the activation pattern of the goal structure within the Multi-Dimensional Memory Frame at the time of speaking. The activation pattern within the Multi-Dimensional Memory Frame, triggered by one's own utterance and the listener's nonverbal responses, is processed in a timely manner via System 2 mediated by C-resonance. This process checks whether there is any discrepancy between this pattern and the activation pattern of the goal structure within the Multi-Dimensional Memory Frame that was activated during speech preparation. When discrepancies are detected, the goal and method are reselected using the activation pattern of the goal structure, which is updated sequentially with utterance via System-2-Before-Event-Mode. The utterance is then continued via System-1-Before-Event-Mode (red arrow).

IV. Discussion

A. Analysis of Conversations Involving Speaker Alternation Based on the Four Processing Modes of MHP/RT

Section III-C2 focused on a single speech event and used Figure 3 to explain the basic processes of PCM and memory of the listener and speaker. Actual conversations are executed by linking together these basic processes. Let E_N denote the utterance termination event for conversation participant B in the basic process shown in Figure 3. We will then examine the relationship between E_N and the events that occurred leading up to it.

1) Smooth Conversation Mode: In the smooth conversation mode, the only event that can be consciously recognized is the end of an utterance. Figure 5 illustrates how the conversation progresses, focusing on the four processing modes of MHP/RT. The utterance termination events indicated by "$\Rightarrow$" for the conversation participant A shown on the left are denoted by E_{N-3}, E_{N-1}, while the utterance termination events indicated by "$\Leftarrow$" for the conversation participant B shown on the right are denoted by E_{N-2}, E_N. The utterance is executed by System-1-Before-Event-Mode, and during the utterance, no correction is performed based on the monitoring results from System 2.

We will denote the four processing modes of MHP/RT associated with the utterance termination event E_i as follows.

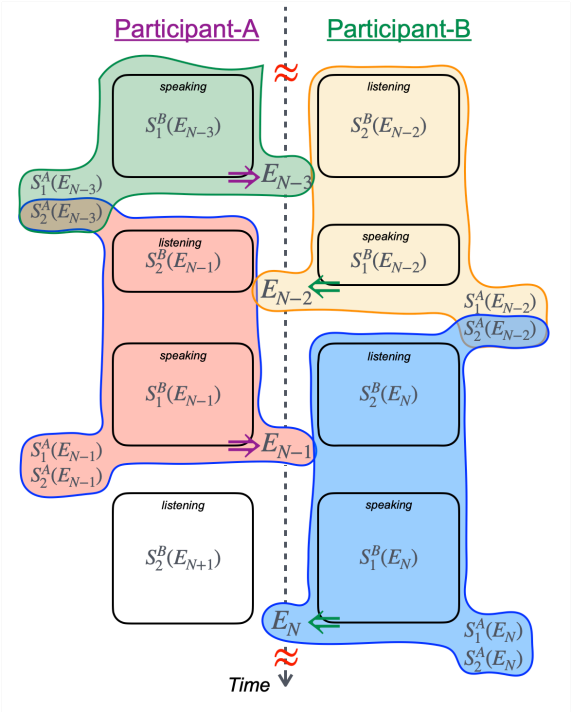


Figure 5. Four processing modes in conversation.

$S_2^B(E_i)$: System-2-Before-Event-Mode
$S_1^B(E_i)$: System-1-Before-Event-Mode
$S_1^A(E_i)$: System-1-After-Event-Mode
$S_2^A(E_i)$: System-2-After-Event-Mode

The parameter β shown in Figure 2 corresponds to the start time of $S_2^B(E_i)$, where $\beta \approx T_i - T_{i-2}$. β' corresponds to the start time of $S_1^B(E_i)$, where $\beta' \approx T_i - T_{i-1}$.

In Figure 5, the portion directly related to participant B's utterance termination event E_N is highlighted with a blue background. Organizing what takes place here in the order of occurrence leads to the following description.

1) While listening to participant A's utterance in $S_1^B(E_{N-1})$, execute one's own utterance preparation in $S_2^B(E_N)$.
2) E_{N-1} occurs when the conversation participant A finishes utterance.
3) Execute one's own utterance in $S_1^B(E_N)$.
4) Finishes one's own utterance, and E_N occurs.
5) Adjust the connection state of Multi-Dimensional Memory Frame nodes unconsciously in $S_1^A(E_N)$.
6) Reflect consciously on E_N in $S_2^A(E_N)$. This will cause changes in the activation patterns in the Multi-Dimensional Memory Frame.

In item 1, the fact that utterance preparation occurs under the conversational partner's utterance is denoted as follows.

$$S_1^B(E_{N-1}) \hookrightarrow S_2^B(E_N)$$

In utterance preparation, the Multi-Dimensional Memory Frame is employed, reflecting the outcome of conscious reflection in $S_2^A(E_{N-2})$ on the event E_{N-2} of one's previous utterance. The Multi-Dimensional Memory Frame of conversation participant B used during utterance preparation reflects the conversation experience E_{N-4}, E_{N-6}, $\cdots$ with conversation participant A up to this point. This is expressed as follows.

$$\cdots,\ S_2^A(E_{N-4}),\ S_2^A(E_{N-2}) \ \mapsto S_2^B(E_N)$$

The act of generating verbal behavior based on the prepared results is denoted as follows.

$$S_2^B(E_N) \mapsto S_1^B(E_N)$$

To summarize the above, the relationship between the event E_N where the conversation participant B terminates the utterance performed in $S_1^B(E_N)$ and the four processing modes of both participants is as follows.

$$\begin{aligned} \text{Participant}_\text{A}\colon\ & S_1^B(E_{N-1}) \ \hookrightarrow \\ \text{Participant}_\text{B}\colon\ & S_2^A(E_{N-2}) \ \mapsto\ S_2^B(E_N) \mapsto S_1^B(E_N) \end{aligned} \tag{1}$$

The relationship between the immediate preceding event, namely event E_{N-1} where the conversation participant A terminates utterance performed in $S_1^B(E_{N-1})$, and the four processing modes of both participants is as follows.

$$\begin{aligned} \text{Participant}_\text{A}\colon\ & S_2^A(E_{N-3}) \ \mapsto\ S_2^B(E_{N-1}) \mapsto S_1^B(E_{N-1}) \\ \text{Participant}_\text{B}\colon\ & S_1^B(E_{N-2}) \ \hookrightarrow \end{aligned} \tag{2}$$

In the smooth conversation mode, (2) connects to (1), generating E_N. A similar relationship applies backward in time. Thus, it becomes clear that a speaker's utterances are influenced by one's own previous utterances and by other's utterances.

2) Intermittent Conversation Mode: When an utterance is being executed in System-1-Before-Event-Mode, System 2 monitors the utterance content in real time to verify that the method is being executed correctly. If no issues are detected, processing by System 1 continues. At this time, the conversation proceeds in the smooth conversation mode. Conversely, when the utterance is judged to have been executed improperly, the method could be reselected and, if necessary, the goal could be re-established. System 2 interrupts the utterance executed in System-1-Before-Event-Mode and, after the cognitive processing in System-1-After-Event-Mode and System-2-After-Event-Mode, the next utterance is executed in System-2-Before-Event-Mode. In this situation, events occur during the speaker's turn without any speaker change. Therefore, the mode in which conversation proceeds in this manner is called the intermittent conversation mode.

Using Figure 3 to explain, conversation participant B begins utterance in $S_1^B(E_N)$ after the turn-taking event E_{N-1}. However, during the process leading up to the next turn-taking event E_N, an interruption by System 2 occurs, and the situation at that moment is made into an event and becomes conscious. Here, "adjustment" (see Section III-C2e) is performed. The reflection of that interruption event is

performed in System-2-After-Event-Mode, and the prepared actions (see Section III-C2b) are modified to reflect the active state of Multi-Dimensional Memory Frame at that time. This involves modifying the utterance goal, which is performed by referencing the goal structure deployed at the Word- and Relation-Multi-Dimensional Memory Frame levels.

The i-th interruption event occurring during the process leading up to event E_N—where conversation participant B finishes speaking and the turn transitions to conversation participant A, as shown in Figure 3—is denoted as $E_{N,i}$. By the time E_N occurs, the following events have taken place:

$$
\begin{aligned}
&\text{Participant}_{\text{B}}\text{:}\\
&S_2^B(E_{N,1}) \rightarrow S_1^B(E_{N,1}) \rightarrow \quad E_{N,1}\\
&\qquad\qquad\qquad\quad \rightarrow \quad S_1^A(E_{N,1}) \quad \rightarrow S_2^A(E_{N,1}) \rightarrow\\
&\qquad\qquad\qquad\quad \vdots\\
&S_2^B(E_{N,i}) \rightarrow S_1^B(E_{N,i}) \rightarrow \quad E_{N,i}\\
&\qquad\qquad\qquad\quad \rightarrow \quad S_1^A(E_{N,i}) \quad \rightarrow S_2^A(E_{N,i}) \rightarrow\\
&\qquad\qquad\quad \cdots \quad \rightarrow \quad \cdots \qquad\quad \rightarrow\\
&\qquad\qquad\quad \cdots \quad \rightarrow \quad E_N
\end{aligned}
$$

When conversation participant B is speaking in the intermittent conversation mode, changes in participant B's utterance goals interfere with listener participant A's ability to consistently execute utterance understanding of participant B's utterance content in $S_2^B(E_{N+1})$. This is because the goal structure activated by conversation participant B in $S_2^B(E_N)$—which was active at the start of the utterance—smoothly connects to the Multi-Dimensional Memory Frame activated by conversation participant A in $S_2^B(E_{N-1})$ while A is processing the utterance. However, as the utterance progresses, this structure gets updated, forcing conversation participant A to follow accordingly.

B. Synchronization in Conversation

1) Synchronization Among Conversation Participants in Conversational Behavior: The utterance content of conversation participant B in event E_N reflects the content of participant B's Multi-Dimensional Memory Frame activated in $S_2^B(E_N)$. This includes an evaluation of the results of participant A's utterance in event E_{N-1} and participant B's own utterance in event E_{N-2}. The content of Speaker A's utterance reflects the content of Speaker A's Multi-Dimensional Memory Frame activated in $S_2^B(E_{N-1})$.

In this way, the Multi-Dimensional Memory Frame possessed by each speaker shifts as the conversation progresses, with the activated domain changing based on the evaluation of both the content of the other's utterances and the results of their own utterances.

Verbal behaviors occur by utilizing the hierarchically structured knowledge network shown in Figure 4. Given this, the necessary condition for a smooth conversation to proceed, where speakers alternately process each other's utterances via their respective System-1-Before-Event-Mode without intervention from System 2, can be summarized as follows.

The activation pattern of Multi-Dimensional Memory Frame via $S_2^B(E_{i-1}) \approx$ The activation pattern of Multi-Dimensional Memory Frame via $S_2^B(E_i)$

This indicates that the activation pattern of one's own Multi-Dimensional Memory Frame that triggers event E_i related to one's own utterance significantly overlaps with the activation pattern of the other's Multi-Dimensional Memory Frame that triggers event E_{i-1} related to the other's utterance. When this condition is met, we can say that conversational behaviors are proceeding synchronously among the conversation participants.

2) Synchronization with VR Systems: The synchronization of conversations involving turn-taking between humans differs in nature from the synchronization that occurs when users interact with VR systems. Dinet et al. [10] identified weak synchronization between the user and the system as the condition for users to interact with multimodal systems with a sense of immersion. Weak synchronization is achieved by designing the specific VR content of the interaction at T_N based on cognitive processing in $S_2^B(E_N),\ S_1^B(E_N),\ S_1^A(E_N),\ S_2^A(E_N)$ regarding the interaction event E_N where the user utilizes the system. To do so, it is necessary to appropriately estimate the propagation of activation within the Multi-Dimensional Memory Frame and the updates to the Multi-Dimensional Memory Frame during the period $[T_N - \beta,\ T_N + \alpha]$.

C. Verbal Behavior and GOMS

This section examines how each element of GOMS relates to conversational behavior. It demonstrates that System 1 and System 2 participating in conversational behavior in a balanced and appropriate proportion is crucial for understanding conversational behavior as it occurs in the real world.

1) The Balance Between Goals and Methods: When we focus on actions that are repeated routinely, these actions manifest based on the GOMS connection structure shown in Figure 4. The individual G, O, M, S shown in Figure 4 are nodes within the Multi-Dimensional Memory Frame corresponding to the knowledge elements necessary to select and execute appropriate actions without continuously referencing the Perceptual-Multi-Dimensional Memory Frame, which is activated through P-resonance with environmental information. Furthermore, action selection and execution must occur in synchronization with an environment whose state changes moment by moment. Therefore, it is reasonable to assume an upper bound exists on their total number. The total number of goals is denoted as $\hat{G}$, total number of methods as $\hat{M}$, total number of selection rules as $\hat{S}$, total number of operators as $\hat{O}$, average depth of the hierarchy as $\bar{N}$, and upper bound on the number of nodes as $\hat{C}$, which is a constant value.

A key feature of the GOMS connection structure is that, due to the finite processing capacity of the brain, either System-2-After-Event-Mode or System-1-After-Event-Mode becomes dominant [9][23]. Depending on the degree of dominance, the following four cases can be considered.

- Case 1: When System-1-After-Event-Mode is dominant, $\hat{M} \gg \hat{G}$ holds.

- Case 2: When System-2-After-Event-Mode is dominant, $\hat{G} \gg \hat{M}$ holds.
- Case 3: When actions occur almost exclusively under System-1-After-Event-Mode, $\hat{M} \gg \hat{G} \sim 0$ holds.
- Case 4: When actions occur almost exclusively under System-2-After-Event-Mode, $\hat{G} \ggg \hat{M} \sim 0$ holds.

2) Understanding Conversational Behavior Through Behavioral Linguistics: The balance between System-1-After-Event-Mode- and System-2-After-Event-Mode-dominance changes depending on the range of communities that the individual is directly and indirectly involved in during their life.

Case 1 corresponds to a smooth conversation mode where one speaker dominates the conversation. The prerequisite for establishing this mode is that the participants share a set of conversational methods. In such behavioral ecology, possessing a set of methods specialized for the situations encountered allows for a perfectly smooth life. Therefore, $\hat{M} \gg \hat{G}$ holds. Since methods, which are cognitive elements, are executed unconsciously, actions driven by System 1 become predominant. In Case 2, each speaker's turn is short, requiring frequent speaker changes to maintain mutual understanding and continue the conversation. When a group consists of both direct communities and indirect societies and/or communication occurs through structural language, System-2-After-Event-Mode becomes the dominant behavioral ecology, resulting in $\hat{G} \gg \hat{M}$. In conversation, reasoning using knowledge stored in the Word- and Relation-Multi-Dimensional Memory Frame have an important role. Flexible adaptation to diverse and changing circumstances is made possible by allocating resources to deliberate System-2-Before-Event-Mode utilizing the goal structures.

Case 1 and Case 2 correspond to conversational behavior in the real-world scenario depicted in Figure 3. The above understanding of these cases was made possible by incorporating in detail how the dual-process—comprising System 1 and System 2, central concepts in behavioral economics—relates to conversational behavior. Therefore, the approach to understanding the verbal behavior demonstrated in this study can be termed behavioral linguistics.

Cases 3 and 4 correspond to ways of understanding that differ from the understanding of conversational behavior proposed in this study. Case 3 involves situations where the goal is extremely limited (such as simply maintaining a conversation). This reduces to stimulus-response behavior that can be executed without cognitive processing, such as producing utterances that can respond to the other person's speech. This aligns with the understanding of verbal behavior based on Skinner's behavioral psychology. Case 4 represents a situation where the methods become extremely limited. The goals determine the details of the utterance. The sequence length of operators leading to the method that specifies the sequence of words to be uttered is short. Consequently, utterances become possible through a large number of goals and their combinations. This aligns with the Chomskyan understanding of linguistic behavior, which posits that goals are symbols and that linguistic actions arise through the manipulation of these symbols.

V. Conclusion and Future Work

This paper proposed a behavioral linguistics theory to explain real-time language generation in everyday conversation. This approach was based on the MHP/RT [6][7] and the dual-process theory of cognition [4][8][15]. The analysis modeled fluent conversation as a PCM cycle alternating between the speaker and the listener, describing in detail how the Multi-Dimensional Memory Frame and GOMS are used for understanding, preparing, and producing utterances. The study also differentiated between fluid and intermittent conversation modes and concluded that this approach, integrating behavioral psychology and behavioral economics, offers a richer framework than models based solely on Skinner [1] or Chomsky [11].

From a theoretical point of view, our paper argued that real-time linguistic behavior cannot be fully captured by static grammatical models or by purely associative accounts of verbal behavior. Instead, it requires a dynamic framework that explains how linguistic choices emerge from moment-to-moment cognitive constraints, environmental cues, and interactive demands. By positioning language production within the PCM cycle, the theory emphasized that utterances are not pre-constructed entities retrieved whole from memory but are assembled progressively through rapid iterations of perception, interpretation, planning, and articulation.

Furthermore, the proposed model highlighted the role of prediction and anticipation in conversation. Speakers and listeners continuously forecast each other's intentions, adapt to turn-taking cues, and adjust their linguistic formulations based on cognitive load and situational incentives. This predictive loop is shaped not only by linguistic competence but also by heuristics, biases, and cost–benefit evaluations, central themes in behavioral economics. As a result, language use is portrayed as an activity controlled by bounded rationality, optimized under real-time processing constraints rather than idealized grammatical rules.

Finally, with our paper, we suggested that a behavioral-linguistic theory grounded in cognitive architecture offers a more comprehensive explanation of conversational competence. It bridged gaps between psycholinguistics, cognitive psychology, and behavioral science, providing a unified account of how humans understand and produce language in everyday interaction. When we understand that speech is built moment-by-moment using limited attention and working memory, we become more aware of why misunderstandings happen. This can help people speak more clearly, listen more actively, and manage turn-taking more smoothly in conversations.

From an applied perspective, several implications can be drawn for our daily lives. Research to realize these implications must be conducted in the future:

(a) Improved learning and teaching. A theory that explains how language is processed in real time can improve language teaching. Teachers can design exercises that match how the brain naturally organizes and retrieves language, making learning more intuitive and efficient;

(b) Reduced communication stress. Knowing that hesitations, pauses, or "ums" are natural results of cognitive processing—not signs of incompetence—can help people feel less anxious when speaking. This is especially helpful for public speaking, second-language use, or social anxiety;
(c) Better HCI. If we understand how humans generate language under time pressure, we can design voice assistants, chatbots, and AI systems that interact more naturally. The theory can guide systems to adapt to human pacing, prediction patterns, and conversational rhythms;
(d) More effective teamwork and decision-making. In workplaces, communication failures are often cognitive failures. Understanding how System 1 and System 2 influence what we say can help people catch biases, avoid rushed judgments, and communicate more thoughtfully in meetings or negotiations;
(e) Insights for therapy and mental health. Speech disruptions often reflect cognitive overload, stress, or emotional pressure. A behavioral model of real-time language can help psychologists better understand how anxiety, ADHD, or trauma affect communication—and help people manage these effects;
(f) Conflict prevention and smoother social interactions. Recognizing that people often speak using quick, automatic processing (System 1) can make us more tolerant of minor errors or emotional reactions in others. It encourages patience and gives a more compassionate understanding of how real conversations work.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI (Grant Numbers 19K12246 / 20H04290 / 22K12284 / 23K11334) and the National University Management Reform Promotion Project.

REFERENCES

[1] B. F. Skinner, *Verbal Behavior*. Appleton-Century-Crofts., 1957.
[2] P. N. Chase, D. W. Ellenwood, and G. Madden, "A Behavior Analytic Analogue of Learning to Use Synonyms, Syntax, and Parts of Speech", *The Analysis of Verbal Behavior*, vol. 24, no. 1, pp. 31–54, 2008. DOI: 10.1007/BF03393055[retrieved: February,2026].
[3] H. A. Simon, "Rational choice and the structure of the environment", *Psychological Review*, vol. 63, pp. 129–138, 1956.
[4] D. Kahneman, "A perspective on judgment and choice", *American Psychologist*, vol. 58, no. 9, pp. 697–720, 2003.
[5] M. Kitajima and M. Toyota, "Simulating navigation behaviour based on the architecture model Model Human Processor with Real-Time Constraints (MHP/RT)", *Behaviour & Information Technology*, vol. 31, no. 1, pp. 41–58, 2012. DOI: 10.1080/0144929X.2011.602427[retrieved:February,2026].
[6] M. Kitajima and M. Toyota, "Decision-making and action selection in Two Minds: An analysis based on Model Human Processor with Realtime Constraints (MHP/RT)", *Biologically Inspired Cognitive Architectures*, vol. 5, pp. 82–93, 2013, ISSN: 2212-683X. DOI: http://dx.doi.org/10.1016/j.bica.2013.05.003[retrieved:February,2026].
[7] M. Kitajima, *Memory and Action Selection in Human-Machine Interaction*. Wiley-ISTE, 2016, ISBN: 9781848219274.
[8] J. S. B. T. Evans, "Dual-processing accounts of reasoning, judgment, and social cognition", *Annual Review of Psychology*, vol. 59, no. Volume 59, 2008, pp. 255–278, 2008, ISSN: 1545-2085. DOI: https://doi.org/10.1146/annurev.psych.59.103006.093629[retrieved:February,2026].
[9] M. Kitajima, M. Toyota, J. Dinet, and K. T. Nakahira, "Transforming Conscious Goals into Unconscious Actions in Real-world Interactions: Real-world Use of Behavioral Ecological Memes via GOMS", *International Journal On Advances in Intelligent Systems*, vol. 18, no. 3 & 4, pp. 173–186, 2025.
[10] J. Dinet and M. Kitajima, "Immersive interfaces for engagement and learning: Cognitive implications", in *Proceedings of the 2015 Virtual Reality International Conference*, ser. VRIC '18, Laval, France: ACM, 2018, 18/04:1–18/04:8, ISBN: 978-1-4503-3313-9. DOI: 10.1145/3234253.3234301[retrieved:February, 2026].
[11] N. Chomsky, *Language*, vol. 35, no. 1, pp. 26–58, 1959.
[12] Y. Barnes-Holmes, S. C. Hayes, D. Barnes-Holmes, and B. Roche, "Relational frame theory: A post-skinnerian account of human language and cognition", in *Advances in Child Development and Behavior*, ser. Advances in Child Development and Behavior, H. W. Reese and R. Kail, Eds., vol. 28, JAI, 2002, pp. 101–138. DOI: https://doi.org/10.1016/S0065-2407(02)80063-5[retrieved:February,2026].
[13] H. Simon, "A Behavioral Model of Rational Choice", *The Quarterly Journal of Economics*, vol. 69, no. 1, pp. 99–118, 1955.
[14] H. A. Simon, *Models of man; social and rational.* Oxford, England: Wiley, 1957.
[15] D. Kahneman, *Thinking, Fast and Slow*. New York, NY: Farrar, Straus and Giroux, 2011.
[16] K. E. Stanovich and R. F. West, "Individual differences in reasoning: Implications for the rationality debate?", *Behavioral and Brain Sciences*, vol. 23, no. 5, pp. 645–665, 2000. DOI: 10.1017/s0140525x00003435[retrieved:February,2026].
[17] S. C. Levinson and F. Torreira, "Timing in turn-taking and its implications for processing models of language", *Frontiers in Psychology*, vol. Volume 6 - 2015, 2015, ISSN: 1664-1078. DOI: 10.3389/fpsyg.2015.00731[retrieved:February,2026].
[18] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation", in 4, vol. 50, Linguistic Society of America, 1974, pp. 696–735.
[19] S. K. Card, T. P. Moran, and A. Newell, *The Psychology of Human-Computer Interaction*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1983.
[20] M. Kitajima et al., "Basic Senses and Their Implications for Immersive Virtual Reality Design", in *AIVR 2024 : The First International Conference on Artificial Intelligence and Immersive Virtual Reality*, 2024, pp. 31–38.
[21] M. Kitajima, M. Toyota, and K. T. Nakahira, "Addressing the Symbol Grounding Problem in VR", in *AIVR 2025 : The Second International Conference on Artificial Intelligence and Immersive Virtual Reality*, 2025, pp. 56–62.
[22] M. Kitajima, M. Toyota, and K. T. Nakahira, "Why the Symbol Grounding Problem Matters in Virtual Reality: A Meme-Focused Solution Based on the Model Human Processor with Real-Time Constraints", *International Journal On Advances in Intelligent Systems*, vol. 18, no. 3 & 4, pp. 162–172, 2025.
[23] M. Kitajima, M. Toyota, J. Dinet, and K. T. Nakahira, "Implementation of Structured Memes into Behavioral Ecology via GOMS", in *COGNITIVE 2025 : The Seventeenth International Conference on Advanced Cognitive Technologies and Applications*, 2025, pp. 6–16.

Behaviour Modeling of Virtual Autonomous Driving Agent Using Voice Command in Risky Scenarios

Velichko Minev
Department Of Computer Systems and Technologies
Technical University of Sofia, Branch Plovdiv,
Plovdiv, Bulgaria
Email: vilim2001@abv.bg

Dilyana Budakova
Department Of Computer Systems and Technologies
Technical University of Sofia, Branch Plovdiv
Plovdiv, Bulgaria
Email: dilyana_budakova@tu-plovdiv.bg

Abstract— **The present paper describes behavior modeling of a multimodal virtual agent–automobile, used in an urban environment under conditions of reduced visibility. In situations in which the virtual agent cannot decide how to act, it receives voice commands from a human assistant. The agent reacts to the voice instructions, reasoning over them, interpreting, and eventually carrying them out. It operates in a hybrid control mode, which includes autonomy with the possibility for a human voice intervention. The architecture of the agent is presented in the paper. Experimental evaluation of the effectiveness of recognition and interpretation of voice commands by a system, based on a Large Language Model (LLM) in risky scenarios with reduced visibility, has been made and described. The results show that the agent's behavior in a risky environment improves as a result of receiving and executing voice commands from a remote operator-assistant. Further research possibilities into behavior modeling of autonomous virtual agents, related to integration of Virtual Reality and Multimodal Large Language Models, are also discussed.**

Keywords-behaviour; modelling; agents; voice commands; tele-assistance; risky scenarios; visibility.

I. INTRODUCTION

Autonomous, driverless electric automobiles, such as Waymo [1], Zoox robot taxis [2], AutoX's self-driving grocery delivery vehicles [3], and Tesla's Full Self-Driving (Supervised) [4], combine inspiring innovative technologies, aiming to provide a safe, environmentally friendly and accessible urban mobility environment for people.

However, there are risky scenarios, such as traffic jams, dense fog, smoke, fire, destruction or rescue operations, in which the automated system cannot independently decide on action. Various approaches, in which a human assistant can remotely give commands to the system, are applied in these cases. The system can be switched to a mode of manual or remote control on the side of a tele-operator, or it can be operated semi-autonomously. The human tele-operator can remotely perform the driving tasks in whole or in part.

This paper describes a developed model of a multimodal virtual autonomous agent–automobile, traveling along a given route in a dynamic urban environment under conditions of reduced visibility due to smoke and fog. The agent–automobile operates in a hybrid control mode, which includes both autonomy and the possibility for a remote intervention by a tele-assistant by means of voice commands. The agent's architecture integrates visual perception, recognition of dynamic objects (people and other cars), interpretation, reasoning, and execution of voice commands via Whisper and Large Language Model (LLM).

An approach has been proposed and experimentally investigated in which voice commands are used to assist the virtual agent-automobile in its driving task when, due to reduced visibility, it cannot decide what action to take. The experiments show that voice commands help reduce uncertainty, and when implemented, the agent's behavior in a risky environment improves. This has been confirmed experimentally in dangerous scenarios by measuring the virtual agent's reaction time and the number of successful passages through various obstacles.

The model combines several modern research directions: multimodal autonomous agents; human-machine interaction through voice advice; 3D modeled virtual environment as an experimental platform; voice control in case of low visibility (fog or smoke); voice control, implemented using Whisper and Large Language Model.

The paper is structured as follows: Section 2 reviews existing solutions for human-assisted autonomous driving. Section 3 details the architecture of the proposed multimodal autonomous agent. Section 4 describes the experimental setup, including the technologies used, the modeling of the virtual environment, and the specific test scenarios. Section 5 presents and analyzes the experimental results, while Section 6 provides conclusions, generalizations, and outlines future trends for autonomous agent development.

II. SUPPORTING THE AUTONOMOUS AGENTS DRIVERS BY HUMAN TELE-ASSISTANTS IN SITUATIONS, IN WHICH THE AGENTS CANNOT MAKE A DECISION INDEPENDENTLY

Conducting repeated tests of autonomous driving systems in risky scenarios is dangerous, expensive, and virtually impossible. Realistic test platforms, based on virtual reality and 3D modeling, are used for these purposes. Such a platform is presented in [5]. Other examples include: Carcraft software developed by Google's Waymo automated driving team; AirSim system for autopilot vehicle testing at Microsoft [6]; Apollo virtual driving platform, created by Baidu Apollo [7].

According to manufacturers and scientists, at this stage, the Autonomous Driving Systems (ADS) may encounter situations in which they cannot make an independent

 ISBN: 978-1-68558-375-0

decision. In these cases, a human being is needed to solve the problem remotely [8].

One solution is Full Self-Driving (Supervised) Systems [4], [9]. They require a human driver to actively monitor the traffic situation during the journey and his/her minimal intervention. The driver can simultaneously see the real situation on the road and watch the road situation as it is perceived by the autonomous system on a screen. Thus, he/she is in a position to react appropriately when registering a discrepancy.

When critical situations occur, a solution package can be provided for autonomous automobile fleets [10]. This package includes support for tele-operations [4], [10], [11], by means of which the fleets are monitored remotely, and real-time information is provided, allowing the operators to offer help when needed.

There is a wide variety of approaches to the ways of solving problems encountered by an autonomous automobile. They are classified according to their complexity in [8], and a taxonomy for Remote Human Input Systems (RHIS) is presented, as well as a Dynamic Driving Task (DDT), where remote assistance and remote driving may be required. Here are examples of such approaches: detecting an object or event, sending information, guiding along a path, and others.

According to [12], [13], innovative models for remote control of autonomous vehicles are needed. A Survey on Tele-operation Concepts for Automated Vehicles is presented in [14]. In [12], the authors explore the construction of a command language as a first step for designing a Tele-assistance user interface. A tablet and command buttons are used for each of the commands for this purpose. When a button is pressed, a command is transmitted to the autonomous vehicle and the vehicle executes it.

According to [15], the integration of LLM into autonomous driving systems will improve the natural language user interface. The autonomous driving systems will explain everything they see and do during a journey. This is supposed to create comfort, trust, and a feeling of security.

LLMs will also help with processing big data (on the scale of the internet data), as well as with managing complex, multi-step scenarios, requiring higher-level reasoning. LLMs will be integrated into planning systems, which will improve understanding of the context of the monitored situation and the corresponding decision-making for a given context [15]. Comparison of LLM-based Autonomous Driving Systems, as well as Comparison of Multimodal Large Language Model-based (MLLM-based) Autonomous Driving Systems, is made in [15].

With the introduction of a video encoder, MLLM systems can directly process visual information from driving scenarios and implement multimodal reasoning. Recent trends in science show that Generative Artificial Intelligence for autonomous driving is approaching the field of embodied AI, such as robotics. This helps to develop vision-language-action (VLA) models.

Since the risky situations are numerous, the need for tele-operation and especially tele-assistance by a human operator is increasingly seen as extremely important. The low latency of 5G networks and the high reliability of wireless communication channels allow to construct “vehicle control towers” where several human assistants can control many vehicles. Challenges related to achieving safety and cybersecurity of such a system are discussed in [16].

III. ARCHITECTURE OF A MULTIMODAL AUTONOMOUS AGENT-VEHICLE

The considered trends in developing remote-control models for autonomous vehicles and the existing guidelines for creating a tele-assistance user interface justify the conduction of a study to assess the effectiveness of using LLM-based recognition systems, reasoning, and executing voice commands, given, e.g., by means of Whisper + LLM in various risky scenarios. What is important is this system's resistance to noise, latency, and the number of misinterpretations of the given commands at critical moments.

This paper examines the degree of effectiveness of using voice commands for controlling automated driving systems in risky scenarios.

The architecture of the multimodal autonomous agent-driver is presented in Figure 1.

The main blocks in this architecture are as follows:

- Multimodal input, supporting: Unity Camera for recognizing the environment, the street, other pedestrians, and vehicles; Unity Radar and Unity Raycast system for determining the distance to dynamic and static objects; Unity Raycast system to model the composition of fog and smoke; Whisper model for automatic recognition of the voice commands, given by a remote tele-assistant.
- A Cognitive module for reasoning and for deciding whether to drive autonomously or to execute voice commands, coming from a remote human assistant. The module includes Multimodal Large Language Modules (MLLM), Whisper model, algorithms for Deep Learning (DL), and for Reinforcement Learning (RL), as well as for Reinforcement Learning with Human Feedback, and Rules for choosing the type of driving control (autonomous or following voice commands from a tele-assistant).
- A module for implementing the movement of the vehicle, controlled by the autonomous agent-driver. It includes the capabilities provided by the multiplatform environment Unity, such as a navigation system, radar models, raycast, and a keyboard.
- Whisper is a state-of-the-art model for Automatic Speech Recognition (ASR) and speech translation [17]. Built as an encoder-decoder model, Whisper processes audio input and extracts a text command, passed to FastAPI. An asynchronous speech activity detection mechanism is used for this

purpose, which records and analyzes the audio signal.

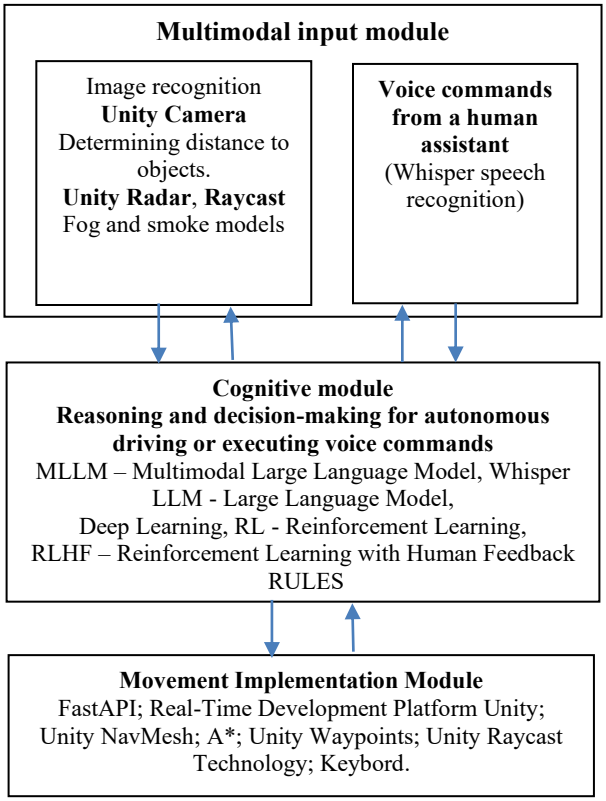


Figure 1. Architecture of a multimodal autonomous agent-driver.

FastAPI [18] is a web framework for building RESTful API interfaces in Python. It provides efficient and asynchronous communication and is therefore suitable for systems processing data in real time. In this project, FastAPI provides the connection between Whisper and the Real-Time Development Platform Unity. It ensures fast and reliable transmission of text commands received through speech recognition. After processing the command, FastAPI returns a confirmation of its successful execution or an error message (e.g., "command accepted" or "command unrecognized"). This ensures clear and interactive communication within the system.

The multiplatform environment for development and 3D simulations, Unity [19], [20], was used to model the realistic virtual scene and the behavior model of the autonomous vehicle. For the agent to drive the automobile intelligently and autonomously in a complex 3D environment, the concept of a navigation mesh (NavMesh) was used. NavMesh allows agents to move smoothly and continuously within defined walkable areas without the need to follow predefined paths. The built-in A* (A-star) algorithm is used to find the shortest or most efficient path through the network of traversable polygons.

Impassable areas are automatically avoided by the virtual agent-automobile by using the NavMesh Agent component. In the context of the system, the commands received from FastAPI (e.g., "go to the intersection") are translated into target coordinates in the virtual environment. These coordinates are then fed to the NavMesh Agent. It autonomously calculates and follows the required path.

The proposed architecture integrates a suite of state-of-the-art technologies to ensure robust collaboration between the human and the agent. Whisper was selected for its resilient encoder-decoder architecture, which provides high-fidelity speech transcription even under the stressful conditions of tele-assistance. The resulting text is processed by Large Language Models (LLM) and Multimodal LLMs (MLLM), which serve as the system's 'cognitive core' by fusing linguistic commands with the visual environmental context to perform high-level reasoning. Reinforcement Learning from Human Feedback (RLHF) is incorporated as a fine-tuning mechanism to align the agent's decision-making with human expectations and safety standards. Finally, the entire framework is deployed within the Unity 3D platform, which provides a high-fidelity, physics-based simulation environment. This allows for the safe testing of edge-case scenarios, such as reduced visibility and obstacle avoidance, prior to real-world implementation.

The information flow begins with the real-time processing of environmental sensor data to assess visibility levels. If visibility falls below the 6-meter threshold, the system triggers a transition to tele-assisted mode, where the Whisper-based module captures audio input. The LLM-based cognitive core then parses this input to extract actionable instructions, which are finally converted into specific vehicle control commands (e.g., steering or braking) within the Unity simulation environment. The system uses a 6-meter visibility threshold as its quantitative uncertainty criterion.

IV. Modeling the Scene, the Agents, and the Scenarios for the Experiments

To conduct the experiments, a realistic and lively urban environment was modeled. An asset [21] was imported, which contained pre-arranged blocks of buildings and a road network. This allows the generation of an extensive and detailed map, on which the virtual autonomous automobile can travel.

Figure 2 shows a top view of the modeled urban environment and shows the route for conducting experimental tests by means of colored lines and numbers. The movement of the autonomous automobile starts at point 1, passes sequentially through points 2 and 3, and ends at point 4. The route from point 1 to point 2 is marked with a red line. The route from point 2 to point 3 is shown with a green line, and the line from point 3 to point 4 is blue.

The experimental route is designed to simulate critical situations in an urban environment. It includes sequential left and right turns, as well as a critical obstacle - a stationary large-scale truck positioned within the driving lane.

The objective is to demonstrate that voice commands from the tele-assistant can prevent a collision with an object that the vehicle's sensors fail to detect in time due to reduced visibility. In this manner, voice commands, such as 'Overtake from the left' serve as a high-level semantic control.

The arrival time of the autonomous automobile at each point on the route was measured by a script, implementing a timer.

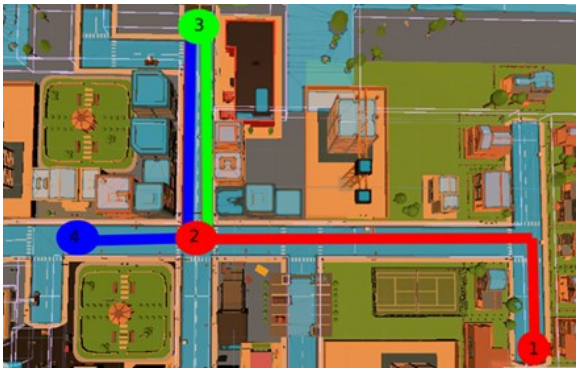


Figure 2. Model of an urban environment. Route for conducting experiments marked with colored lines and numbering.

To model traffic and achieve a realistic atmosphere, 3D models of the virtual autonomous automobile, of pedestrians, and other vehicles were implemented. To achieve the fog effect, the Unity “Fog” function was used, and its parameters were adjusted.

Raycast technology, supported by Unity, was used to simulate reduced visibility in foggy conditions. The visual range of the virtual autonomous agent-automobile is reduced by shortening the length of the rays used to detect obstacles. This leads to the delayed detection of obstacle objects and, accordingly, to a delayed reaction of the agent-automobile to avoid them. This impacts the reliability and safety of autonomous driving under conditions of low visibility.

In this study, the term 'scenarios' refers to a combination of external factors and the methods used to control the vehicle. The external factors to which the autonomous vehicle must adapt include the degree of fog and specific road conditions. The control methods include autonomous mode and voice command tele-assisted mode.

Consequently, the experimental design is structured as a 2*3 matrix: Control Modes (Autonomous vs. Voice-Assisted) and External Factors, which include full visibility, varying degrees of limited visibility, and the emergence of a complex situation (the obstacle). This structure enables a rigorous evaluation of how human intervention mitigates risks when the agent's autonomous capabilities are challenged by both environmental and situational factors.

Consequently, the experiments comprise the following scenarios:

• Monitoring autonomous movement of an autonomous automobile in a dynamic urban environment, without and with the use of voice command tele-assistance.

According to the following conditions and problems:

- Clear and sunny weather.
- Limited visibility due to dense fog.
- An event of a complex situation, which the autonomous agent cannot resolve independently.

The following characteristics were considered:

- The average time from giving a voice command to its execution.
- The number of successfully recognized and executed commands in the event of voice control.
- The total time for the different scenarios for completing the whole route from point 1 to the end point 4.
- Whether situations arise in autonomous driving in which voice commands are needed in order to continue driving (for example, a stuck autonomous vehicle).

The tele-assistant continuously monitors the autonomous vehicle and provides support through voice commands. A confidence threshold has been established, which is determined by visibility levels in foggy conditions. In clear weather, the vehicle prioritizes autonomous operation, maintaining a high level of confidence. However, when visibility drops to six (6) meters or less (quantitative uncertainty criterion), the agent’s confidence level decreases, and it begins to execute all voice commands provided. In critical situations, the autonomous vehicle is unable to proceed without the assistance of the tele-assistant. At this stage of the system's implementation, the agent places full trust in the tele-assistant. Future work could introduce a safety module to verify command-context consistency.

V. Experimental Results and Discussion

The results of the experiments reflect the arithmetic mean of the time it takes for the automobile to get from the starting point to the end point of the route and are presented in seconds.

Table 1 presents the results of the experiments conducted for autonomous vehicle control under conditions of full visibility of the order of 20 meters and when modeling reduced visibility of the order of eight (8), six (6), and four (4) meters. The reported time for the autonomous automobile to travel the entire route is given in the second column of Table 1. The last third column indicates whether and under what visibility conditions the need to assist autonomous driving with voice commands arises.

The results show that when visibility is reduced to six (6) meters or fewer, for example, four (4) meters, there is a need for driving assistance with voice commands. Under these risky conditions, the time to complete the entire route increases.

Table 2 presents the results of the experiments, conducted with an autonomous agent-automobile, receiving voice commands under conditions of both full visibility and reduced visibility, caused by fog. The first column of Table 2 gives the meteorological conditions modeled for each group of experiments.

The result in the form of the response time of the autonomous vehicle when receiving a voice command is given in the second column in seconds. The third column shows the time for completion of the route with the help of voice commands. The last column indicates the number of unrecognized voice commands.

TABLE I. RESULTS FROM THE BEHAVIOR OF A VIRTUAL AGENT-VEHICLE DURING A CHANGE IN VISIBILITY. TIME TO TRAVEL THE GIVEN ROUTE. NEED FOR USE OF VOICE COMMANDS

Meteorological conditions. Degree of Visibility simulated with raycast	Time to travel on the given route	Need for voice commands to continue the agent's movement along the route
Sunny Rays Visibility – 20 m.	80,36 sec.	no
Foggy Rays Visibility – 8 m.	80,47 sec.	no
Foggy Rays Visibility – 6 m.	107,72 sec.	yes
Foggy Rays Visibility – 6 m.	109,23 sec.	yes
Foggy Rays Visibility – 4 m.	117,07 sec.	yes
Foggy Rays Visibility т – 4 m.	123,63 sec.	yes

TABLE II RESULTS OF THE BEHAVIOR OF A VIRTUAL AGENT-VEHICLE DURING DRIVING UNDER DIFFERENT METEOROLOGICAL CONDITIONS. REACTION TIME FOR EXECUTING VOICE COMMANDS. NUMBER OF UNRECOGNIZED VOICE COMMANDS

Meteorological conditions	Voice command response time	Route travel time	Number of unrecognized commands
Sunny	2,66 sec.	105,95 sec.	1 num.
Sunny	2,68 sec.	96,76 sec.	0 num.
Sunny	2,64 sec.	98,77 sec.	0 num.
Foggy	2,69 sec.	109,76 sec.	0 num.
Foggy	3,00 sec.	123,85 sec.	2 num.
Foggy	3,21 sec.	117,85 sec.	0 num.

The results show a high extent of recognition of voice commands. Only two commands were not recognized in foggy conditions and one in sunny weather.

These commands include 'Turn left' and 'Overtake the truck from the right.' The failure to recognize voice commands can be attributed to changes in the tele-assistant's intonation. Stress in critical situations can impact the tele-operator's psychological state, potentially leading to alterations in the volume, speed, and intonation of the command, as well as its specific wording. Additionally, technical issues, such as audio signal interruptions may occur.

If the voice command is not recognized, the human assistant must repeat it. The autonomous vehicle's response time to voice commands is of the order of three (3) seconds.

Therefore, when using voice commands and communication between an autonomous vehicle and a human operator, this delay must be considered to effectively assist driving and avoid accidents.

The effectiveness of facilitating driving by voice commands depends on their timely delivery. The experiments show that the agent-driver may collide with an obstacle or stop if the voice command (such as turn left, go around, turn right) is not given in time.

Under conditions of simulated dense fog, a change in the behavior of the agent-automobile was observed. It executed the voice commands of the human assistant, and driving effectiveness was improved.

In risky conditions, such as reduced visibility, when there was no communication with a human assistant, it was observed that the autonomous automobile did not detect obstacles or detected them too late. This led to collisions or to situations in which the agent could not decide to take an action and stopped moving.

There were some cases, in which to continue along the route proved impossible without additional intervention by means of voice commands, such as "go around to the right" or "avoid left". As a result of the experiments, a conclusion was drawn that in the event of reduced visibility and when complex situations arise, the agent-driver needs assistance in the form of voice commands, short and clear, and given in a timely manner.

The speed of voice command transmission, i.e., the lack of a delay, is extremely important. The voice command recognition software is critical, as it must ensure correct recognition. And, finally, the speed of the autonomous vehicle's centralized electronic supercomputer architecture, which can execute the assigned commands in a timely manner, is also of great significance.

VI. CONCLUSIONS AND FUTURE WORK

This paper investigates the effectiveness of using voice commands in modeling the behavior of a multimodal autonomous virtual agent-automobile, driving in a risky environment of reduced visibility caused by fog or smoke.

The architecture of the agent is proposed, and an improvement in its behavior when executing voice commands in critical scenarios with reduced visibility is shown.

To conduct the research, a 3D scene and a model of a virtual autonomous agent-automobile are implemented. Unity Real-Time Development Platform and the Whisper model for automatic speech recognition are used. A fog model and varying degrees of visibility are included. The obtained results, concerning the behavior of the autonomous automobile, the speed of the system and the effectiveness of its operation both when using voice commands in tele-assistance and without their use, are discussed.

The results show that in risky scenarios, the use of tele-assistance with voice commands for Autonomous Driver Systems is both necessary and very effective.

Voice commands are used as high-level semantic control and reduce uncertainty, rather than replacing low-level vehicle control.

An advantage is the use of a large language model, such as the Unity Whisper model, to implement communication with voice commands. LLMs allow for the use of natural language expressions in communication. The commands can

correctly be interpreted by the autonomous vehicle without requiring humans to learn a special set of commands.

Future iterations of the system will address potential LLM risks, such as hallucinations, by incorporating a cross-verification safety module to validate commands against real-time sensor data. Additionally, a robust fail-safe protocol is planned to ensure a controlled vehicle stop in the event of connectivity or model failure.

Another question to be cleared is at what point the agent-vehicle should seek help from a human tele-assistant or from other passengers and vehicles around.

We believe that it is worth exploring the need and form of communication to be maintained between vehicles in a risky environment in one and the same area, in close proximity to each other. It is also of importance to create a balance, allowing for discussing a particular situation by neighboring vehicles on the road and for avoiding distraction and inattention.

We also place emphasis on researching the need to ensure protection and cybersecurity of Tele-operation of Connected and Automated Vehicles.

The paper also considers the latest inspiring technologies introduced in the implementation of ADS and the opportunities for further research, created by Virtual Reality, Generative Artificial Intelligence, LLM, and MLLM.

ACKNOWLEDGMENT

The authors gratefully acknowledge the financial support provided by the Project No.: BG-RRP-2.004-0005 “Improving research capacity and quality for international recognition and sustainability of Technical University of Sofia”; National Recovery and Sustainability Plan, BG-RRP-2.004 - Creating a network of research universities in Bulgaria.

REFERENCES

[1] “Waymo - Self-Driving Cars - Autonomous Vehicles - Ride-Hail,” Waymo. Accessed: Dec. 30, 2025. [Online]. Available: https://waymo.com/index/ [retrieved: February, 2026]

[2] K. Korosec, “Zoox becomes fourth company to land driverless testing permit in California,” TechCrunch. Accessed: Dec. 30, 2025. [Online]. Available: https://techcrunch.com/2020/09/18/zoox-becomes-fourth-company-to-land-driverless-testing-permit-in-california/ [retrieved: February 2026]

[3] “Journey – AutoX.” Accessed: Dec. 30, 2025. [Online]. Available: https://www.autox.ai/en/journey.html [retrieved: February 2026]

[4] “Full Self-Driving (Supervised) | Tesla.” Accessed: Dec. 30, 2025. [Online]. Available: https://www.tesla.com/fsd [retrieved: February, 2026]

[5] S. Yao, J. Zhang, Z. Hu, Y. Wang, and X. Zhou, “Autonomous-driving vehicle test technology based on virtual reality,” The Journal of Engineering, vol. 2018, no. 16, pp. 1768–1771, 2018, doi: 10.1049/joe.2018.8303.

[6] S. Shah, D. Dey, C. Lovett, and A. Kapoor, “AirSim: High-Fidelity Visual and Physical Simulation for Autonomous Vehicles,” in Field and Service Robotics, M. Hutter and R. Siegwart, Eds., Cham: Springer International Publishing, 2018, pp. 621–635. doi: 10.1007/978-3-319-67361-5_40.

[7] Q. Liu, “Baidu Turnip Fast Running Self-driving Car Industry Development Status and the Discussion of Existing Problems,” Journal of Education, Humanities and Social Sciences, vol. 48, pp. 51–56, Mar. 2025, doi: 10.54097/732kma16.

[8] D. Bogdoll, S. Orf, L. Töttel, and J. M. Zöllner, “Taxonomy and Survey on Remote Human Input Systems for Driving Automation Systems,” in Advances in Information and Communication, K. Arai, Ed., Cham: Springer International Publishing, 2022, pp. 94–108. doi: 10.1007/978-3-030-98015-3_6.

[9] T. Zhang, “Toward Automated Vehicle Teleoperation: Vision, Opportunities, and Challenges,” IEEE Internet of Things Journal, vol. 7, no. 12, pp. 11347–11354, Dec. 2020, doi: 10.1109/JIOT.2020.3028766.

[10] “Nuro Toolkit,” Nuro. Accessed: Dec. 30, 2025. [Online]. Available: https://www.nuro.ai/nuro-toolkit, [retrieved: February, 2026]

[11] S. Lu, R. Zhong, and W. Shi, “Teleoperation Technologies for Enhancing Connected and Autonomous Vehicles,” in 2022 IEEE 19th International Conference on Mobile Ad Hoc and Smart Systems (MASS), Oct. 2022, pp. 435–443. doi: 10.1109/MASS56207.2022.00068.

[12] F. Tener and J. Lanir, “Devising a High-Level Command Language for the Teleoperation of Autonomous Vehicles,” International Journal of Human–Computer Interaction, vol. 41, no. 9, pp. 5299–5315, May 2025, doi: 10.1080/10447318.2024.2359224.

[13] C. Kettwich, A. Schrank, and M. Oehl, “Teleoperation of Highly Automated Vehicles in Public Transport: User-Centered Design of a Human-Machine Interface for Remote-Operation and Its Expert Usability Evaluation,” Multimodal Technologies and Interaction, vol. 5, no. 5, p. 26, May 2021, doi: 10.3390/mti5050026.

[14] D. Majstorovic, S. Hoffmann, F. Pfab, A. Schimpe, M.-M. Wolf, and F. Diermeyer, Survey on Teleoperation Concepts for Automated Vehicles. 2022. doi: 10.48550/arXiv.2208.08876.

[15] Y. Wang et. al., Generative AI for Autonomous Driving: Frontiers and Opportunities. 2025. doi: 10.48550/arXiv.2505.08854.

[16] F. J. Jiang, J. Mårtensson, and K. H. Johansson, “Safe Teleoperation of Connected and Automated Vehicles,” in Cyber-Physical-Human Systems: Fundamentals and Applications, IEEE, 2023, pp. 251–272. doi: 10.1002/9781119857433.ch10.

[17] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust Speech Recognition via Large-Scale Weak Supervision,” Dec. 06, 2022, arXiv: arXiv:2212.04356. doi: 10.48550/arXiv.2212.04356.

[18] FastAPI - Introduction. (17:38:21+00:00). GeeksforGeeks. https://www.geeksforgeeks.org/python/fastapi-introduction/ [retrieved: February, 2026]

[19] N. F. Hutchins, L. Hook, W. Friedel, and Z. Kirkendoll, “Use of Unity in Scientific Simulation and Modeling for Research and Education,” Jan. 2017, Accessed: Jan. 02, 2026. [Online]. Available: https://www.academia.edu/115586868/Use_of_Unity_in_Scientific_Simulation_and_Modeling_for_Research_and_Education [retrieved: February 2026]

[20] U. Technologies, “Unity - Manual: Unity 6.1 User Manual.” Accessed: May 02, 2025. [Online]. Available: https://docs.unity3d.com/6000.1/Documentation/Manual/UnityManual.html [retrieved: February, 2026]

[21] “The Best Assets for Game Making | Unity Asset Store.” Accessed: Dec. 30, 2025. [Online]. Available: https://assetstore.unity.com/ [retrieved: February, 2026]

A Reference Architecture for Pro-adaptive Cognitive Assistive Technology

Sebastian Hauscheid[1,*], Sarah Büscher[2], Sinan Yavuz[1], Jordan Schneider[3], Michał Stolarz[3], André Frank Krause[2], Robin Grashof[1], Oviya Rajavel[3], Swathy Satheesan Cheruvalath[3], Teena Chakkalayil Hassan[3], Christian Ressel[2], Nele Wild-Wall[2], Edwin Naroska[1], Thomas Nitsche[1]

Centre for Assistive Technologies Rhine-Ruhr

[1]Hochschule Niederrhein University of Applied Sciences, Krefeld, Germany

[2]Rhine-Waal University of Applied Sciences, Kamp-Lintfort, Germany

[3]Bonn-Rhein-Sieg University of Applied Sciences, Sankt Augustin, Germany

[*]e-mail: sebastian.hauscheid@gmail.com

{sinan.yavuz | robin.grashof | edwin.naroska | thomas.nitsche}@hs-niederrhein.de

{jordan.schneider | michal.stolarz | teena.hassan}@h-brs.de

{oviya.rajavel | swathy.cheruvalath}@smail.inf.h-brs.de

{sarah.buescher | andrefrank.krause | christian.ressel | nele.wild-wall}
@hochschule-rhein-waal.de

***Abstract*—Cognitive Assistive Technology (CAT) provides support for individuals in specific activities and in everyday life. It is used to monitor parameters and activities, act as a reminder, or to inform users of threatening scenarios. The development of CAT for people with cognitive impairments poses a particular challenge due to requirements regarding data protection, acceptance, mobility, and consideration of disease-specific limitations. In neurodegenerative diseases, such as Parkinson's disease or dementia, the limitations and thus motor- as well as non-motor-symptoms increase, resulting in assistive systems that are only able to adapt to specific needs at a given moment and therefore may no longer remain suitable over time. Instead, systems with higher levels of adaptability are required, eventually culminating in Pro-adaptive Cognitive Assistive Technology (Pro-CAT). To better enable the design of highly adaptive assistive systems, this paper presents a reference architecture for Pro-CAT, which serves as a foundation for systems that predict the progression of the user's condition and can adapt the assistance to the needs and limitations accordingly. If the assistance is provided proactively, Pro-CAT falls into the category of adaptive agents.**

Keywords-pro-adaptive cognitive assistive technology; reference architecture; privacy-by-design; levels of adaptability of assistive technology; adaptive systems.

I. Introduction

Population aging is accelerating worldwide along with a rising prevalence of disability, creating substantial challenges for individuals and societies alike. According to projections from the United Nations, there were approximately 727 million people aged 65 years or older in 2020, representing 9.3% of the global population. This number is expected to more than double to over 1.5 billion by 2050, corresponding to 16% of the global population [1]. In addition, the World Health Organization estimates that approximately 15% of the global population lives with some form of disability, of whom 2%–4% experience significant difficulties in functioning, and that prevalence rises sharply with age [2]. By closing the gap between intrinsic capacity and environmental barriers, an Assistive Technology (AT) can play an important role in promoting functional ability in such individuals [3] with neurodegenerative disorders like dementia or Parkinson's disease, entailing cognitive symptoms [4].

ATs encompass a broad range of technical interventions that aim to support daily functioning across various contexts, such as home care [5][6][7], workplace support [8][9][10] and rehabilitation [11][12]. Within this landscape, ATs that anticipate changes in user capabilities and adapt themselves accordingly, known as Pro-adaptive Cognitive Assistive Technology (Pro-CAT) [13], have the potential to align assistance with individual requirements. This highlights the need for structured design approaches to integrate predictive adaptation into ATs.

Reference architectures have been found to reduce reduplicated effort, lower risk through implementing proven solutions and encourage compatibility between systems [14][15][16]. In this paper, we present a reference architecture for a Pro-CAT and introduce the essential components to meet both functional and non-functional requirements. Most importantly, we contribute to the field of the reference architectures for ATs by providing an architecture that can not only support models at run-time but also focuses and provides a framework for introducing and maintaining pro-adaptive components. Moreover, we also provide privacy components that can serve as a guideline for meeting privacy regulations that might be required from the system. The aim of this work is to provide a basis for future Pro-CAT to identify potential risks and corresponding mitigation measures. In addition, the reference architecture enables the comparison of ATs and supports modular changes and easy adaptation of software and hardware components.

The paper is organized as follows: Section II reviews related work and summarizes recent reference architectures for ATs. Section III introduces the concept of Pro-CAT and discusses different levels of adaptability. Section IV presents the results of this work, including the identified requirements and the proposed reference architecture. Section V discusses

and evaluates the proposed architecture, including limitations. Finally, Section VI concludes the paper and outlines directions for future work.

II. Related work

In [17], the authors propose a software reference architecture for ATs. Their contribution is an architecture that supports models at run-time for providing assistance and is achieved through (i) investigating multiple literature sources of functional requirements for the ATs, (ii) defining a set of components and inter-component connections building a reference architecture and (iii) validating the proposed architecture in two scenarios. The scenarios include assistance for elderly people where the system acquires knowledge about the user's behavior and helps by making suggestions once any daily activity is found problematic. The additional scenario is a CAT for operators in manufacturing, where a digital twin supports operators in the production processes.

A reference architecture targeted towards Active and Healthy Aging (AHA) has been proposed in [18]. Here, the aim is to improve the quality of life for elderly living at their homes through an Internet of Things (IoT) system, thus it is an advancement in the field of Ambient Assisted Living (AAL) [19]. As opposed to a very generic reference architecture proposed in [17], this is domain specific; however, it supports integration of different platforms, technologies, and standards for deploying a large-scale system for AHA. Here, the evaluation process is rather focused on testing and assuring interoperability of the modules and components that are part of the deployed example architecture rather than a theoretical analysis of the applicability of the suggested reference architecture to different scenarios (as in [17]).

ATs can also include robots, such as socially assistive robots [20], which can provide support in education or healthcare, and an approach for a reference architecture helping in designing those is proposed in [21]. The authors present a template for the organization of the software components so that the development of the features enabling robot's personalizable social interaction is facilitated. Unlike the aforementioned reference architectures, this one can be only applied while developing robots and was instantiated and evaluated on a social robot supporting people with dementia. Finally, [22] presented a software reference architecture for IoT-based healthcare applications focused on monitoring patients and the surrounding environment. It is designed for broader scenarios than [18] and is more specific than [17]. Moreover, unlike [17], the reference architecture in [22] has been designed based not only on functional but also nonfunctional requirements, such as security or availability. The evaluation is conducted by instantiating the reference architecture in the form of the platform integrating patients and caregivers to facilitate fast reactions in case of critical situations.

The reference architectures discussed in [17][18][20] and [22] do not accommodate the design principles of Pro-CAT [13]. Moreover, some architectures [17][20] overlook critical security aspects, raising ethical concerns especially when developing CAT for vulnerable groups.

III. Pro-Adaptive Cognitive Assistive Technology (Pro-CAT)

ATs have increasingly incorporated adaptive and personalized mechanisms to better accommodate individual user needs. Over time, this has led to a broad spectrum of capabilities and configuration options [23]. However, most existing systems remain primarily reactive, adapting only to the current context or explicitly configured parameters. Such systems are inflexible to adapt automatically to changing user requirements.

Pro-CAT extends this paradigm by explicitly considering temporal dynamics and longitudinal trends. Instead of responding only to the present situation, pro-adaptive systems anticipate future changes in user abilities and proactively adjust assistance strategies to maintain effectiveness, usability, independence, and user acceptance over time.

A. Levels of Adaptability of Assistive Technology

Analogous to the concept of technology-readiness levels [24] and levels of driving autonomy (level 0: no automation to level 5: full driving automation, [25]), we define different *levels of adaptability* of Assistive Technology (AT):

1) Static AT: Assistance is available to the degree defined by the manufacturer of the AT.
2) Adaptive AT: The extent of assistance is adapted based on current sensor data readings.
3) Personalizable AT: Users can manually adapt the AT to their specific needs and requirements.
4) Auto-individualizing AT: The AT adapts to the user automatically based on past observed behavior and sensor data.
5) Pro-adaptive AT: The functionalities of items 2 to 4 are combined. Moreover, the AT adapts to meet the user's (future) needs based on anticipated conditions that are derived from past and current sensor data (e.g., aging or disease related changes in the user's capabilities).

Personalizable AT (level 3) is well established and can be found in consumer products like cars. For example, users can adjust the steering control sensitivity, engine response curves ('sport mode') or cabin air conditioning (for a review, see [26]). Systems with a level 4 adaptability were conceptualized in the context of Advanced Driver Assistance (ADAS), suggesting a modular decomposition for the next generation of personalized ADAS and human machine interaction, which can be expected to continuously adapt in interaction with the driver [27]. For example, researchers from Toyota and the Georgia Institute of Technology developed a data-driven approach to personalized autonomous driving [28]. This approach automatically adapts the driving style of autonomous vehicles based on observed driving data of an individual user. Another example of a level 4 adaptive AT, in which assistance is adjusted based on past sensor data, was proposed in [29]. Here, the authors present a behavior model that enables a socially assistive robot to adapt to the user based on the past interaction. The model

is implemented as a deep neural network that is trained on multiple input modalities and outputs actions to be performed by the robot.

B. Concept and Definition of Pro-Adaptive Cognitive Assistive Technology

While existing AT aims to address a user's immediate needs, the concept of Pro-CAT, as introduced in [13], is intended to have a broader scope. Various aspects are to be taken into account in order to offer users the most effective support feasible: Pro-CAT aims to continuously evaluate its users' situational data, e.g., concerning health or learning progress, the context in which the users are situated, and their abilities, thereby learning over a longer period of use how the user's state, aging process, and condition develop and change. On this basis, a prediction of how the user's condition may develop in a given amount of time can be calculated and the CAT may provide the user with the best possible support at the current stage as well as simultaneously prepare them for the expected course. Due to the predictive functionality, in the event of deteriorating abilities, compensation options can be practiced with the user at an early stage to well prepare them before their condition worsens. In certain cases, these measures may help maintain abilities and increase acceptance over time. The system operates in such a way that tasks which users are still able to master independently are not performed for them or assisted strongly. Likewise, as skills are acquired and the condition improves, the level of support is reduced accordingly. Providing more support than needed at a given time may lead to negative effects, such as dependence on the AT or a deterioration in the user's abilities, as skills may be unlearned. In Pro-CAT, the concept of a 'human digital twin' [13] may be applied, which aims to display, model and continuously adjust all parameters and data relating to a person on the basis of regular measurements, and to predict the further course of development.

Following our concept paper [13], the interplay of three core components can enable pro-adaptive systems (see Figure 1): 1. a prognostics module that predicts changes in relevant abilities based on competence- and progression models, 2. a cognitive state estimation module that assesses the user's current state, and 3. a context module inferring user intentions. On top, continuous learning may be applied to update the models over time to deliver highly individualized assistance.

The key advantages of Pro-CAT include sustained personalization, improved long-term usability, and the ability to proactively adjust assistance strategies in response to predictable changes in user capabilities.

IV. Pro-CAT Architecture

This section briefly presents functional and non-functional requirements of a Pro-CAT. Furthermore, the proposed reference architecture for a Pro-CAT is described in detail, including the components and functionality.

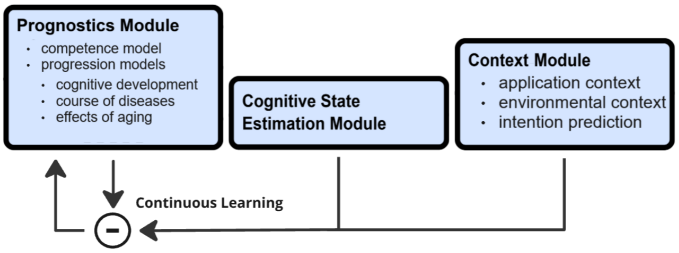


Figure 1. Components enabling Pro-CAT: User data and environmental information is processed by three modules: 1. a prognostics module predicting changes in relevant abilities, 2. a cognitive state estimator assessing the user's current state, and 3. a context module inferring user intentions. Continuous learning updates the models over time to deliver individualized assistance.

A. Identified Requirements of Assistive Technologies

For the development of AT, several functional and non-functional requirements must be met. The detection of disease-related symptoms, in addition to the monitoring of parameters, such as pulse, activity, and noise, is a significant requirement, especially for people with cognitive impairment or neurodegenerative disorders, such as Parkinson's disease or dementia, since disease progression worsens over time and there is no cure. An example of this is the detection of tremor or freezing of gait in people with Parkinson's disease, or the measurement of activities as well as focusing time for a specific task in patients with Attention Deficit Hyperactivity Disorder (ADHD). The local storage of detected symptoms can support long-term tracking of, e.g., a neurodegenerative disease like Parkinson, and, in addition, be used to train an AI-based digital twin of the patient in order to model the disease course. This can help and support physicians as well as patients in medication therapy. In addition to the detection of disease-related symptoms, the detection of falls and reminder functions for appointments, medication, or fluid intake are also essential components of cognitive AT especially for elderly people. The preliminary detection of these events can reduce falls and thus the risk of further injuries. In this case, real-time capability and optimized sensor data processing are also important requirements. To actively support the motor or non-motor functions of the patient, the AT should provide assistance based on the capabilities of the patient, which are reduced over time in neurodegenerative disorders. In this case, the AT adapts and provides assistance according to the specific user's needs to avoid over-assistance [30].

As symptom-related constraints may limit interaction, usability must be given. These constraints include motor impairments (e.g., tremor or bradykinesia), cognitive limitations (e.g., reduced attention, memory, or executive functioning) or sensory limitations (e.g., impaired sense of vision or hearing). Consequently, user interfaces and output components should be adaptive and configurable, presenting only the information necessary for the current situation, while respecting individual preferences and disease-related requirements (e.g., larger targets, reduced interaction steps, reminder timing, modality selection) [30][31]. Moreover, the reference architecture must

support transparency and explainability of AI-driven behavior, enabling users to understand why recommendations, prompts, or adaptations occur, thereby fostering trust, perceived control, and safe use. The architecture should enable local computing to ensure low latency, robust operation during connectivity disruptions and privacy-preserving processing of sensitive data (e.g., on-device inference and local data minimization). While local autonomy is critical, the system must retain cloud integration capability to support external resources and services, specifically including interoperability with cloud-based services. Significant offerings are updates, model life-cycle management, cross-device synchronization, or optional analytics, without making core functionality dependent on continuous cloud access. Finally, the architecture must be expandable, supporting incremental integration of additional components and ecosystems (especially smart home systems) through modular design and standards-based interoperability. To prevent the collection of data from third parties in AT that use environmental sensors, additional security measures are required alongside local data storage, anonymization, and data encryption. These measures may include rules for specific scenarios, such as visits from family members, caregivers, or physicians, during which data acquisition from certain sensors is disabled or completely stopped. To this end, these rules can be configured and activated using a physical switch called privacy switch in our reference architecture.

B. Components

For the selection of suitable components, various reference architectures for ATs were analyzed, and frequently occurring components were adopted. In addition, we identified and specified components required to enable **Pro-Adaptivity** of CAT. The result is shown in Table I, which lists all components together with their respective functionality.

TABLE I. COMPONENTS AND THEIR TASKS

ID	Component	Task
C1	Anonymization (Security)	Removes or masks sensitive data to protect privacy.
C2	Encryption / Decryption	Encrypts data for secure transmission and decrypts incoming data.
C3	Privacy Switch	Physical switch to interrupt data acquisition or transmission.
C4	Privacy Mode Handler	Controls privacy modes, stops data flows if necessary, or restricts system access.
C5	Storage Manager	Manages stored data and its organization.
C6	Physical Storage	Physical storage of data (e.g., database, file system).
C7	Data Management	Organizes, manages, and provides data within the system.
C8	Actuators	Execute physical actions as defined by the system.
C9	Actuator Control Unit	Translates system decisions into control commands for actuators.
C10	Acquisition Devices	Devices for data acquisition, e.g., sensors or cameras.
C11	AI Management	Manages AI models, including loading, configuring, or switching.
C12	Pro-Adaptivity	Adapts the system to changing environments or user needs.
C13	System Administration / System Log	Monitors and logs system activities, supports maintenance and error analysis.
C14	External Communication	Interface for communication with external systems or services.
C15	User Interface	Visual/interactive interface for user interaction and result presentation.

C. Proposed Reference Architecture

The reference architecture (shown in Figure 2) developed in this paper follows a modular, privacy-by-design approach for Pro-CATs. It is designed to support secure data acquisition, adaptive AI-based processing, and context-aware actuation while maintaining a clear separation of concerns between data handling, decision-making, and system control. Privacy enforcement, Pro-Adaptivity, and AI management are treated as first-class architectural elements, enabling both local and externally supported intelligence under configurable privacy constraints.

Interaction with the system occurs through three primary channels:

(i) the `Privacy Switch`, which defines which data may be processed, stored, or forwarded,

(ii) the `User Interface`, through which services are selected and information is displayed,

(iii) the `Acquisition Devices` (e.g., wearables or environmental sensors), which are responsible for data acquisition.

The `Privacy Mode Handler` receives signals from the `Privacy Switch` and enforces the corresponding privacy and operational policies throughout the system. Depending on its configuration, it regulates the processing and transmission of data, particularly with respect to `Data Management` and `External Communication`.

In Figure 2, the process of allowing or denying external communication is visualized using a diamond-shaped decision element. This explicitly highlights that external communication is not implicit but subject to a configuration-dependent decision by the `Privacy Mode Handler`, ensuring privacy-by-design and a clear separation of concerns.

Sensor data collected by the `Acquisition Devices` are first anonymized within `Anonymization (Security)` before being forwarded to `Data Management`.

`Data Management` acts as the central integration component of the architecture. It ingests data from multiple sources, performs preprocessing and structuring, and routes information to downstream components, such as `AI Management`, the `Storage Manager`, `Pro-Adaptivity`, or `External Communication`. Relevant information may also be fed back to the `User Interface`.

The `AI Management` component comprises two main functional areas. The first functional area (i), the *AI Component*, is responsible for executing the deployed AI models (inference). In this process, AI models are loaded from the `Storage`

Manager and applied to the incoming data in order to generate predictions or classifications.

The second functional area is the *Model Optimizer*. This component is responsible for the training or optimization of AI models based on feedback from Pro-Adaptivity. In this context, optimization may include updating model parameters (weights) or adjusting model configurations to improve system performance.

The resulting updated model versions or parameters are then stored again in the Storage Manager. Depending on the application scenario, the updated models can subsequently be loaded and executed by the *AI Component*.

Pro-Adaptivity processes inputs from AI Management and Data Management, requests additional external data when required via Data Management and transmits its results to the User Interface and the Actuator Control Unit to trigger actions if necessary. Its key capability is self-adaptation: it analyzes variations in inputs or system states, adjusts internal parameters or model weights accordingly, and returns optimized values to AI Management, allowing the system to dynamically adapt to new conditions and evolving user needs.

Beyond technical adaptation, Pro-Adaptivity explicitly addresses cognitive aspects of assistance. By anticipating changes in user capabilities, the system can gradually adjust assistance intensity, avoid abrupt behavioral changes, and prevent negative psychological effects, such as over-reliance or loss of agency. This cognitive perspective is essential for long-term acceptance and distinguishes Pro-CAT from purely reactive adaptive systems.

The User Interface visualizes system and status information received from Pro-Adaptivity, the Actuator Control Unit, System Administration/System Log, and Data Management. In addition, it forwards user inputs to Data Management.

The Actuator Control Unit translates high-level decisions into device-specific commands for the Actuators, which execute actions in the physical environment.

The Storage Manager coordinates data flows to persistent storage, performs encryption and decryption of sensitive information via Encryption/Decryption, and interacts bidirectionally with Physical Storage as well as with AI Management for storing models and weights.

External Communication encapsulates interaction with external systems, such as cloud-based AI services. Access is governed by the Privacy Mode Handler, while data and results are exchanged bidirectionally with Data Management. It also connects to external AI services that provide additional models or analytical capabilities.

Finally, System Administration/System Log monitors operational states, collects event logs, and supports traceability and auditing of security- and privacy-relevant processes.

V. Discussion | Evaluation

This paper proposes a modular, privacy-by-design reference architecture for Pro-CAT [13], which is evaluated in this section from an architectural and quality-attribute–driven perspective. The architecture deliberately separates concerns between (i) privacy control and enforcement, (ii) data acquisition and management, (iii) AI model operation and optimization, and (iv) pro-adaptive decision-making and actuation. This separation constitutes a key architectural design decision and directly supports maintainability, extensibility, and reuse: individual components can be replaced or specialized (e.g., different AI models, storage backends, or communication mechanisms) without requiring structural changes to the overall system.

A. Privacy-by-Design

A key architectural decision is to treat privacy enforcement as a first-class concern. The combination of the Privacy Switch and the Privacy Mode Handler enables explicit user control and systematic policy enforcement across internal processing and external communication. This is particularly relevant for sensitive domains, such as healthcare and education, where data minimization and controllable data flows are critical for acceptance and compliance. In addition, the explicit integration of Anonymization (Security) and Encryption/Decryption provides a clear architectural pathway to implement privacy-by-design in a consistent and reusable manner, rather than relying on ad-hoc measures.

B. Pro-Adaptivity

The architecture emphasizes Pro-Adaptivity as a distinct capability, implemented through the interaction between AI Management and Data Management. In contrast to purely reactive adaptation based only on the current context, pro-adaptive behavior requires longitudinal signals, temporal modeling, and feedback loops for continuously adjusting parameters, models, or assistance strategies. The architecture supports this by (i) enabling data integration from diverse sources, (ii) allowing model updates and persistence over time, and (iii) supporting closed-loop feedback from observed outcomes to model optimization. As a result, Pro-Adaptivity is treated as an explicit architectural capability rather than an implicit consequence of AI-based processing.

C. Limitations

- **Conceptual validation:** The reference architecture is derived from literature and design reasoning and has not yet been validated through a longitudinal deployment or controlled user study.
- **Operationalization of Pro-Adaptivity:** The paper outlines pro-adaptive mechanisms at an architectural level; the concrete selection of prognostic models, update strategies, and calibration procedures depends on the application domain and remains to be specified.

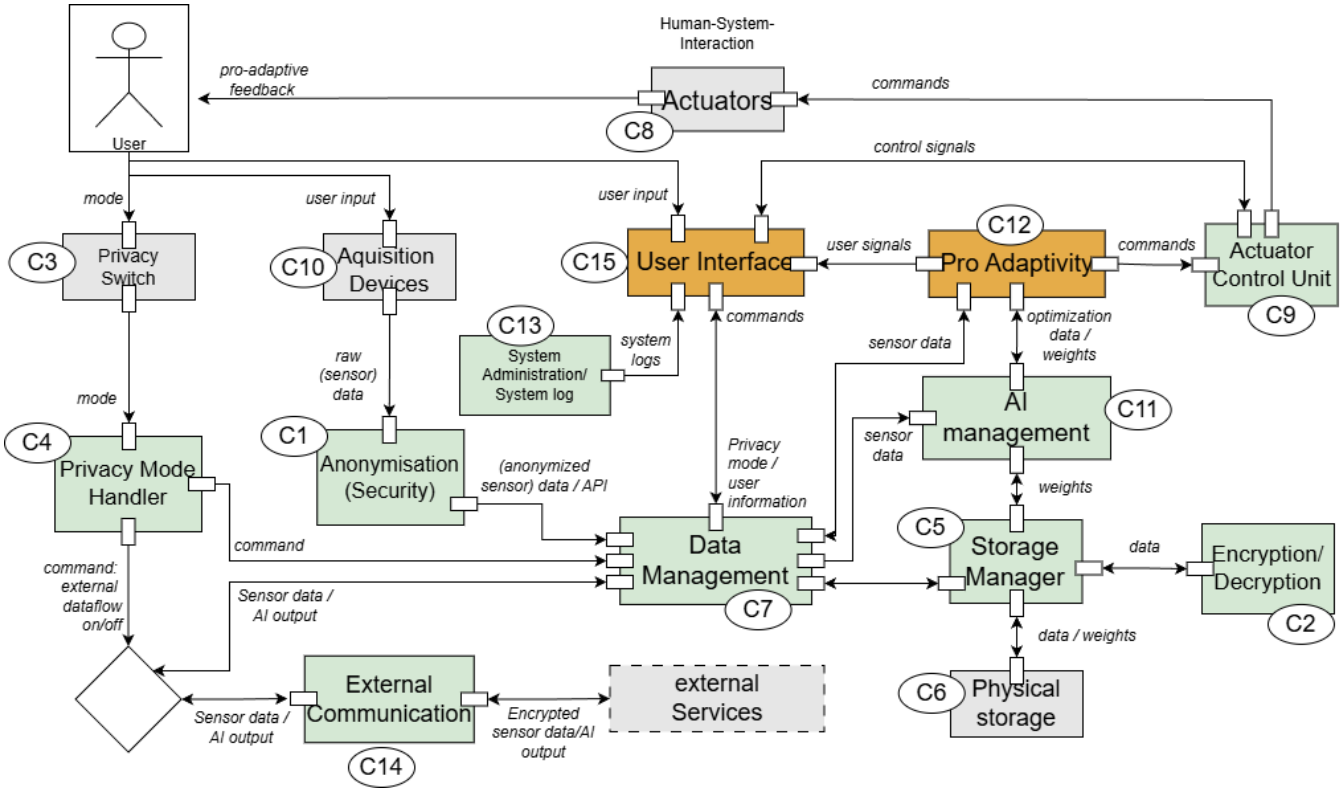


Figure 2. Overview of the reference architecture for Pro-CAT, which depicts the interaction between privacy control, data management, adaptive AI management, and actuation.

VI. Conclusion and Future Work

The proposed architecture addresses a key limitation of existing ATs by explicitly supporting longitudinal adaptation to predictable changes in user capabilities, such as those resulting from aging, progressive disease, recovery, learning, or skill acquisition. In contrast to purely reactive or manually configurable systems, the architecture enables pro-adaptive behavior by integrating continuous data acquisition, AI-based analysis, and feedback-driven model refinement within a clearly structured architectural framework.

A central contribution of this work is the explicit treatment of privacy and Pro-Adaptivity as first-class architectural concerns. The architecture supports transparent data governance, user control, and compliance with privacy regulations while maintaining adaptability and system autonomy. The clear separation of concerns between data management, AI management, decision-making, and actuation enhances modularity, reusability, and maintainability, allowing individual components to be replaced or extended without affecting the overall system structure.

Furthermore, the reference architecture provides a common conceptual basis for comparing, designing, and evolving pro-adaptive ATs across different application domains. It supports both local, privacy-preserving operation and optional cloud-based services, thereby balancing robustness, low latency, and extensibility.

We also provide a concept of levels of adaptability of assistive technologies analogous to technology-readiness levels. As such, the architecture contributes not only a technical blueprint but also a structured guideline for future Pro-CAT.

Future work should focus on making the architecture actionable and empirically grounded:

1) **Monitoring, evaluation, and user studies:** Evaluate the architecture through implementation, e.g., in scenarios outlined in this paper (Parkinson support, dementia support, ADHD support, cognitive training), using measurable criteria, such as usability and acceptance, perceived control, adaptation quality, and system robustness. Longitudinal studies are particularly important to evaluate the benefits of Pro-Adaptivity over purely reactive adaptation.
2) **Prognostic module and temporal modeling:** Specify how prognostics and longitudinal user modeling are realized within `Pro-Adaptivity` (or as an explicit submodule), including update triggers, uncertainty management, and personalization strategies.

Overall, the proposed reference architecture provides a structured foundation for engineering pro-adaptive cognitive ATs. By making privacy and Pro-Adaptivity explicit architectural concerns, it supports both trustworthy data handling and continuous personalization.

Acknowledgement

This work was supported by the Ministry of Culture and Science of the State of North Rhine-Westphalia as part of the project *Center for Assistive Technology Rhine-Ruhr* (ZAT)

(11/2023 to 10/2026, Grant No. PB22-076A, PB22-076B, PB22-076C, PB22-076D).

Author Contributions. SH, SB, SY, JS, MS, AFK, RG, OR, SSC, NWW and TN contributed to the manuscript. CR, NWW, TCH and EN acquired funding. CR and NWW coordinate the ZAT project.

REFERENCES

[1] Y. Kamiya, N. M. S. Lai, and K. Schmid, "World population ageing 2020 highlights", *United Nations Department of Economic and Social Affairs*, 2020.

[2] W. H. Organization et al., "World report on disability", in *World report on disability*, 2011, pp. 24–24.

[3] W. H. Organization, *World report on ageing and health*. World Health Organization, 2015.

[4] R. Grashof, M. Lipprandt, and B. Breil, "Cognitive assistive technologies for degenerative diseases and related evaluation methods: A scoping review", *GMS Med Inform Biom Epidemiol*, vol. 21, no. 9, 2025. DOI: 10.3205/mibe000281.

[5] R. Al-Shaqi, M. Mourshed, and Y. Rezgui, "Progress in ambient assisted systems for independent living by the elderly", *SpringerPlus*, vol. 5, no. 1, p. 624, 2016. DOI: 10.1186/s40064-016-2272-8.

[6] M. Z. Uddin, W. Khaksar, and J. Torresen, "Ambient sensors for elderly care and independent living: A survey", *Sensors*, vol. 18, no. 7, p. 2027, 2018. DOI: 10.3390/s18072027.

[7] F. Zhou, J. R. Jiao, S. Chen, and D. Zhang, "A case-driven ambient intelligence system for elderly in-home assistance applications", *IEEE Transactions on Systems, Man, and Cybernetics - Part C (Applications and Reviews)*, vol. 41, no. 2, pp. 179–189, 2011. DOI: 10.1109/TSMCC.2010.2052456.

[8] T. Marinaci et al., "An inclusive workplace approach to disability through assistive technologies: A systematic review and thematic analysis of the literature", *Societies*, vol. 13, no. 11, p. 231, 2023. DOI: 10.3390/soc13110231.

[9] J. Wolfartsberger, J. D. H. Haslwanter, and R. Lindorfer, "Perspectives on assistive systems for manual assembly tasks in industry", *Technologies*, vol. 7, no. 1, p. 12, 2019. DOI: 10.3390/technologies7010012.

[10] N. Mandischer et al., "Toward adaptive human–robot collaboration for the inclusion of people with disabilities in manual labor tasks", *Electronics*, vol. 12, no. 5, p. 1118, 2023. DOI: 10.3390/electronics12051118.

[11] A. Guatibonza, L. Solaque, A. Velasco, and L. Peñuela, "Assistive robotics for upper limb physical rehabilitation: A systematic review and future prospects", *Chinese Journal of Mechanical Engineering*, vol. 37, 2024. DOI: 10.1186/s10033-024-01056-y.

[12] A. Nanavati, V. Ranganeni, and M. Cakmak, "Physically assistive robots: A systematic review of mobile and manipulator robots that physically assist people with disabilities", *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 7, pp. 123–147, 2024. DOI: 10.1146/annurev-control-062823-024352.

[13] A. F. Krause, K. Kannen, S. Büscher, C. Ressel, and N. Wild-Wall, "Pro-adaptive cognitive assistive technology: Concept and application in reading support for adhd", in *International Conference on Extended Reality*, Springer, 2025, pp. 255–266.

[14] S. Martínez-Fernández, C. P. Ayala, X. Franch, and H. M. Marques, "Benefits and drawbacks of software reference architectures: A case study", *Information and software technology*, vol. 88, pp. 37–52, 2017.

[15] E. Y. Nakagawa, P. Oliveira Antonino, and M. Becker, "Reference architecture and product line architecture: A subtle but critical difference", in *European conference on software architecture*, Springer, 2011, pp. 207–211.

[16] P. H. Dias Valle, L. Garcés, T. Volpato, S. Martínez-Fernández, and E. Y. Nakagawa, "Towards suitable description of reference architectures", en, *PeerJ Comput. Sci.*, vol. 7, no. e392, e392, Mar. 2021.

[17] J. Michael and V. A. Shekhovtsov, "A model-based reference architecture for complex assistive systems and its application", *Softw. Syst. Model.*, vol. 23, no. 5, pp. 1247–1274, Mar. 2024, ISSN: 1619-1366.

[18] C. I. Valero et al., "Aiotes: Setting the principles for semantic interoperable and modern iot-enabled reference architecture for active and healthy ageing ecosystems", *Computer Communications*, vol. 177, pp. 96–111, 2021.

[19] C. Dobre, C. x. Mavromoustakis, N. Garcia, R. I. Goleva, and G. Mastorakis, *Ambient Assisted Living and Enhanced Living Environments: Principles, Technologies and Control*, 1st. USA: Butterworth-Heinemann, 2016, ISBN: 0128051957.

[20] D. Feil-Seifer and M. Mataric, "Defining socially assistive robotics", in *9th International Conference on Rehabilitation Robotics, 2005. ICORR 2005.*, 2005, pp. 465–468. DOI: 10.1109/ICORR.2005.1501143.

[21] L. Asprino, P. Ciancarini, A. G. Nuzzolese, V. Presutti, and A. Russo, "A reference architecture for social robots", *Journal of Web Semantics*, vol. 72, p. 100 683, 2022. DOI: 10.1016/j.websem.2021.100683.

[22] I. de Morais Barroca Filho, G. S. Aquino Junior, and T. B. Vasconcelos, "Extending and instantiating a software reference architecture for IoT-based healthcare applications", in *International Conference on Computational Science and Its Applications*, Springer, 2019, pp. 203–218.

[23] M. Zallio and T. Ohashi, *The evolution of assistive technology: A literature review of technology developments and applications*, 2022. arXiv: 2201.07152 [cs.HC].

[24] J. C. Mankins, *Technology readiness levels – a white paper*, 1995.

[25] O.-R. A. D. Committee, *Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles (SAE Standard J3016$_2$02104)*. SAE international, 2021. DOI: https://doi.org/10.4271/J3016_202104.

[26] X. Liao et al., "A review of personalization in driving behavior: Dataset, modeling, and validation", *IEEE Transactions on Intelligent Vehicles*, 2024.

[27] M. Hasenjäger, M. Heckmann, and H. Wersing, "A survey of personalization for advanced driver assistance systems", *IEEE Transactions on Intelligent Vehicles*, vol. 5, no. 2, pp. 335–344, 2019.

[28] M. L. Schrum, E. Sumner, M. C. Gombolay, and A. Best, "Maveric: A data-driven approach to personalized autonomous driving", *IEEE Transactions on Robotics*, vol. 40, pp. 1952–1965, 2024.

[29] M. Stolarz, M. Romeo, A. Mitrevski, and P. G. Plöger, "Deep Learning-Based Adaptation of Robot Behaviour for Assistive Robotics", in *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*, 2024, pp. 110–117.

[30] S. Yavuz, R. Grashof, T. Nitsche, B. Breil, and E. Naroska, "Development of a pro-adaptive wrist-worn wearable device for parkinson disease symptoms: Concept and initial approach", *Biomedical Engineering / Biomedizinische Technik*, 2025. DOI: doi:10.1515/bmt-2025-1001.

[31] R. Grashof, S. Yavuz, J. Gräbel, and B. Breil, "Interviews on the user-centered development of pro-adaptive cognitive assistance systems: Needs of alzheimer's patients", *Deutsche Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie, editors. 70. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e. V. (GMDS)*, 2025. DOI: 10.3205/25gmds102.

Towards Individualised Reading Support for Attention-Deficit/Hyperactivity Disorder (ADHD): User-Centred Development of an Adaptive Eye-Tracking-Based Reading Assistance System

Kyra Kannen*
Rhine-Waal University of Applied Sciences
Kamp-Lintfort, Germany
email: kyra.kannen
@hochschule-rhein-waal.de

Sarah Büscher*
Rhine-Waal University of Applied Sciences
Kamp-Lintfort, Germany
email: sarah.buescher
@hochschule-rhein-waal.de

André Frank Krause
Rhine-Waal University of Applied Sciences
Kamp-Lintfort, Germany
email: andrefrank.krause
@hochschule-rhein-waal.de

Hafsa Ashfaque
Rhine-Waal University of Applied Sciences
Kamp-Lintfort, Germany
email: hafsa.ashfaque
@hsrw.org

Christian Ressel
Rhine-Waal University of Applied Sciences
Kamp-Lintfort, Germany
email: christian.ressel
@hochschule-rhein-waal.de

Nele Wild-Wall
Rhine-Waal University of Applied Sciences
Kamp-Lintfort, Germany
email: nele.wild-wall
@hochschule-rhein-waal.de

***Abstract*—People with Attention-Deficit/Hyperactivity Disorder (ADHD) may experience increased effort and variability in reading performance, including higher cognitive load, attentional fluctuations, and differences in eye movement patterns. While prior research has provided valuable insights into reading challenges associated with ADHD, there remains a limited understanding of how effective reading support can be designed, and a lack of empirically informed reading assistance systems developed in collaboration with and tailored to the needs of people with ADHD. Based on prior research, an early-stage reading assistance prototype with text highlighting features was implemented and refined through a participatory workshop with adults with diagnosed or suspected ADHD (N = 7), combining prototype exploration, group discussions, and co-design activities. The findings provide initial design principles for ADHD-specific reading assistance systems and highlight the value of integrating empirical evidence and user perspectives in assistive technology development.**

Keywords-attention deficit; ADHD; assistive technology; user-centred design; eye-tracking.

I. Introduction

People diagnosed with Attention-Deficit/Hyperactivity Disorder frequently encounter challenges when reading long, complex, or non-interesting texts. Difficulties in sustaining attention and impaired oculomotor control are among the key factors contributing to these challenges. Empirical evidence indicates that individuals with ADHD exhibit increased fixation counts and duration, a higher incidence of regressive and vertical saccades, and decreased reading speed relative to neurotypical controls [1]–[5]. Beyond oculomotor differences, constraints in visual perception, including reduced contrast and colour sensitivity, have been documented in this population [6]–[8]. Compounding these difficulties, mind wandering, a prevalent symptom of ADHD [9], has been associated with reduced text comprehension [10], rendering sustained reading particularly demanding for individuals with this condition.

Assistive Systems (AS) have the potential to support people with ADHD while reading. Nevertheless, a review of assistive technologies for people with ADHD does not identify any AS specifically targeting reading [11], with existing systems primarily focusing on task organisation, reminders, or attention training. Research in this area remains limited and predominantly concentrates on samples with dyslexia or reading and spelling difficulties [12]. Only a small number of studies have explored assistive reading approaches for individuals with ADHD, including text-to-speech technologies [13][14] and text highlighting as a means of guiding attention during reading [15]. To date, no effective, scientifically validated assistive reading system exists for this population, and while adaptive assistive reading systems have recently been proposed [16][17], none have been developed or evaluated specifically for individuals with ADHD. Commercial products such as the *Bionic Reading* font [18] have not yet demonstrated improvements in reading performance in either neurotypical readers [19] or those with ADHD. Given the societal importance of reading and the lack of effective support tools, there is a clear need to investigate requirements for assistance that enhances reading comfort and experience for adults with ADHD.

To address this, we developed an early-stage prototype of an eye-tracking-based assistive reading system, grounded in existing research and design principles, and refined through a user-centred, participatory workshop with adults with ADHD.

 ISBN: 978-1-68558-375-0

Exploratory findings indicate that line spacing received the highest overall ratings and font highlighting was most preferred in the eye-tracking condition, that a high degree of personalisation is essential to accommodate the heterogeneous needs of users with ADHD, and that cursor-based interaction may offer a viable and more accessible alternative to eye tracking without compromising usability.

This paper makes the following contributions:

- The design and implementation of an early-stage assistive reading prototype for adults with ADHD, grounded in empirical evidence and existing design principles.
- Exploratory insights from a user-centred, participatory workshop with adults with ADHD, covering usability, user experience, and reading support preferences.
- Initial design principles for ADHD-specific reading assistance systems, derived from workshop findings and structured around concrete design implications.
- Methodological insights for integrating target users into early stages of assistive technology research and development.

This paper is structured as follows: Section II details the development of the prototype and the design of the user-centred workshop conducted to evaluate it. Section III presents the analysis of the data gathered during the workshop, the results of which are subsequently interpreted and discussed in Section IV. Section V concludes the paper and outlines directions for future research.

II. METHODS

This section outlines the iterative development of an early-stage assistive reading prototype and the methodology employed in the subsequent user-centred workshop conducted to evaluate and refine it.

A. Prototype Development

In order to develop a prototype that best supports the defined target group, the following approach was taken: Users read texts displayed on a screen, while their gaze direction is monitored using a remote eye-tracking system. This facilitates the tracking of the reader's gaze position on the screen, enabling the identification of the currently viewed line and word. The corresponding line is highlighted to assist the reader in maintaining focus and preventing skipping to another line. If the expected reading flow from left to right and line by line is maintained, the highlight moves accordingly. A short delay was implemented before the highlight follows an unexpected gaze jump, briefly retaining focus on the previously read position to help users recover their reading position. To ensure accuracy, a short calibration of the eye tracker is performed when the prototype is started. Four highlighting features were implemented, each grounded in prior research: (1) *background highlighting*, in which the currently viewed line receives a coloured background [15]; (2) *font highlighting*, in which the text colour of the focused line and fixated word is changed; (3) *bionic reading*, in which the first half of each word is rendered in bold, applied dynamically to the currently focused line [19][20]; and (4) *blurring*, in which non-focused lines are rendered as a grey shadow using a Gaussian blur [15]. The prototype employs a light beige background to maintain high contrast and reduce visual strain [21][22]. Users can customise font, background, and highlight colours, as well as font size and line spacing, via a cursor-controlled side menu. Figure 1 illustrates the four main assistive reading features.

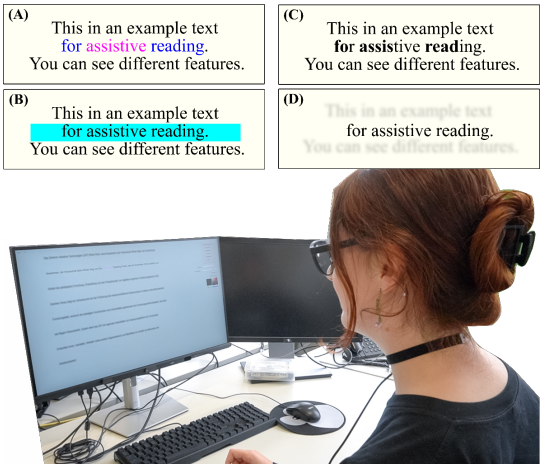


Figure 1. Assistive digital reading features. (A) Font Highlighting, (B) Background Highlighting, (C) Bionic Reading, and (D) Blurring. While reading, gaze can be tracked using eye-tracking glasses, a remote eye tracker or mouse cursor.

B. User-centred Workshop

A participatory workshop evaluated the early-stage prototype, examining user perceptions of attention-aware reading features and collecting preliminary usability and design feedback. The features were assessed using two interaction modalities: the aforementioned eye-tracking-based interaction, and an additional mouse cursor-based interaction in which participants manually moved the cursor to simulate their gaze and trigger text highlighting.

1) Exploratory Research Method: To characterise individual ADHD symptom profiles, all participants completed a demographic questionnaire, including details of ADHD diagnosis (e.g., subtype and current status), and the ADHD Self-Report Scale V1.1 (ASRS-V1.1) [23], a widely used screening for ADHD symptoms. To assess participants' reading difficulties, the Adult Checklist (AC) [24] for dyslexia screening and the Adult Reading History Questionnaire (ARHQ) [25] were administered. System usability and user experience were assessed using the System Usability Scale (SUS) [26] and a custom questionnaire comparing cursor- and eye-tracking navigation, in which participants rated each digital reading feature on perceived helpfulness and ease of reading using a five-point Likert scale.

2) Workshop Procedure: The user-centred workshop was conceptualised as a mixed-method evaluation approach with a scheduled duration of 120 minutes. After informing participants about the research goals and workshop procedure, written informed consent was obtained. Participants received

an expense allowance of 12 € per hour for their participation. First, participants completed the demographic questionnaire, ASRS-V1.1 and AC. Afterwards, a demonstration of the prototype was conducted to ensure that all participants had a shared understanding of the early-stage system. Then, participants were split into two groups. For prototype testing (Group 1), participants individually tested both interaction modalities, rated each visual reading support feature using the self-constructed questionnaire, and completed the SUS. Prior to the group activity (Group 2), participants completed the ARHQ to assess their reading habits, followed by a guided group discussion in which participants noted down their experiences and strategies. Given that many adults with ADHD have developed a range of strategies for coping with the demands of neurotypically shaped societies, the group discussion was used as a method to externalise these often implicit strategies and make them accessible for analysis. Afterwards, the two groups switched activities.

After both groups completed system testing, the subsequent activity focused on co-creation design. All participants received screenshots of the interface they had tested and were asked to (1) criticise the current interface layout and visual design, (2) suggest improvements or changes and (3) comment on alternative design proposals prepared by the researchers. The activity focused on visual design, layout, and interface clarity, rather than feature functionality. Our proposed features include adaptive elements, such as motion tracking that detects high movement, which may be linked to high cognitive demand [27][28] and, therefore, triggers reading assistance or suggestions for a break, alongside text simplification via Large Language Models (LLMs) and a wearable device providing vibration alerts for scheduled appointments to help users manage time during reading.

III. Preliminary findings and insights

A total of 7 participants took part in the workshop (5 female, M_{age}= 27.57 years, SD_{age} = 13.05), with a formal diagnosis of ADHD (n = 4) or a suggested, as yet unconfirmed, diagnosis of ADHD (n = 3). Of the four participants diagnosed with ADHD, two were diagnosed with the inattentive subtype, one with the combined subtype, while the subtype of one participant was unknown. Four subjects were identified as having ADHD based on ASRS-V1.1 screening, suggesting that their symptoms are consistent with ADHD in adults, while the mean dyslexia score, assessed by the AC amounted to 44.14 (SD = 9.72), suggesting only a negligible risk for dyslexia. By contrast, the high mean score of 44.29 on the ARHQ (SD = 10.29) indicates that there are several features of the participants' reading and learning history that are indicative of dyslexia.

A. Descriptive Analysis

The tested prototype achieved a mean SUS score of 81.79 (SD = 12.48), indicating a good overall usability of the prototype. Ratings were similar for eye-tracking (M = 3.41, SD = 0.49) and cursor-based interaction (M = 3.37, SD = 0.42), suggesting no clear preference for either interaction mode. Figure 2 depicts the helpfulness ratings for each feature, differentiated by interaction mode.

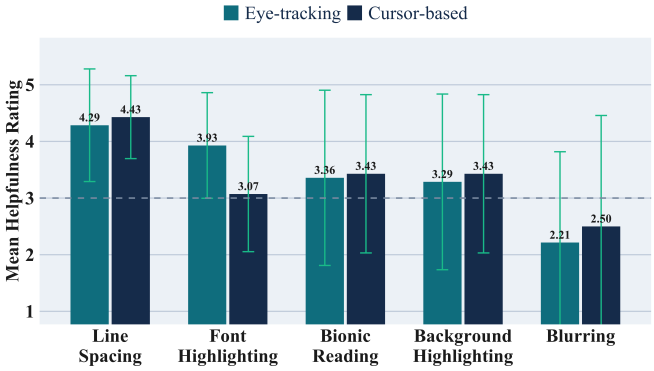


Figure 2. Feature helpfulness ratings by interaction mode. 5-point Likert scale (1 = not helpful, 5 = very helpful). Error bars represent SD.

B. Explorative Qualitative Analysis

The analysis of exploratory qualitative data has revealed a tension between perceived support in reading and the cognitive or physical strain caused by visual aids. While font- and background-highlighting were generally reported to enhance subjective focus and comprehension, users emphasised that the intensity or persistence of visual reading support could increase cognitive load and divert attention from the text. Visual comfort emerged as a key determinant of usability: colour choice, contrast, and personalisation strongly influenced acceptance, with poorly calibrated or overly intense colours described as distracting despite the functional value of the feature.

Moreover, the performance of the prototype further mediated user experience. Delays in eye- or cursor-tracking and lag in visual feedback were consistently described as disruptive to reading flow. This was particularly salient for eye-tracking interactions, where calibration difficulties and latency reduced trust in the system and limited its effectiveness for some participants.

Blurring features produced divergent responses, reflecting individual preferences. Whereas most participants found blurring distracting or uncomfortable, a minority reported that subtle blurring supported line focus. Eye-tracking interaction was valued for enabling hands-free reading and reducing manual effort. However, feedback of eye strain, mild nausea and headaches indicate potential physical costs and highlight the need for optional deployment.

During discussions, participants emphasised the value of features that simplify or summarise text, particularly for academic literature and complex forms. Movement tracking was considered potentially useful if it leverages existing devices, such as smartwatches or smart glasses, rather than additional gadgets. For a potential reminder feature, some participants indicated a preference for visual notifications rather than vibration-based alerts. Text-to-speech functionality was widely

endorsed. Participants also expressed a preference for profiles that store individual visual support settings, minimalistic customisation menus, and features that estimate reading time based on their typical reading speed.

Across all features, participants consistently articulated a need for customisation and feature combinability. Adjustable parameters and layered configurations were seen as essential for balancing support with comfort, underscoring the importance of flexible, user-centred design of assistance systems.

IV. Discussion

Building on existing research on reading challenges experienced by people diagnosed with ADHD, we presented the development of an ADHD-focused reading assistance prototype and reported initial insights from a first user-centred workshop.

Overall, the prototype showed initial indicators of good usability, as evidenced by the high mean SUS score, thus suggesting that the core interaction concept may be accessible and broadly aligned with user expectations. Participants reported a subjective impression of support during reading tasks, providing preliminary evidence for the feasibility of the proposed approach. The results of the user-centred workshop revealed that line spacing received the highest overall ratings. Among the highlighting features, no consistent preference was observed in the cursor-based condition, whereas font highlighting was rated most favourably in the eye-tracking condition. In contrast, the blurring feature received the lowest overall rating, although the high standard deviation hints that it was rated positively by some participants, indicating that highly visually prominent support features may not be equally suitable for all users. This result is in line with research by Shen et al. [15], in which the blurring option had the least positive impact on reading performance. Participants commonly reported that the presence of a single, clearly visible line within the blurring feature caused irritation. While the feature might prevent slippage in the line, it potentially had an adverse effect on reading flow, as the user's visual focus shifts to the next line, but the cursor does not move to this line simultaneously.One potential feature improvement could be to display three lines together, including the line above and below the currently focused line. Font highlighting received the highest ratings, and, in contrast to other features described as 'slow' despite equivalent functionality, this perception did not apply to it. The observed discrepancy could be explained by the feature's capacity to highlight both a full line and the word currently in focus. This may provide users with continuous feedback regarding the location of their gaze, thereby facilitating a more precise comprehension of the system's accuracy and responsiveness. Bionic reading and background highlighting received mixed reviews in both the cursor- and eye-tracking-based versions, with many comments on background highlighting relating to the pre-set colour. This once again underlines the importance of a highly customisable system. The adjustable line spacing has generally received positive feedback, although criticism may be related to limitations of the prototype, as some text was not displayed correctly with wide line spacing.

Interestingly, there was no clear preference regarding prototype interaction method (mouse cursor vs. eye tracking). Nevertheless, in all conditions, cursor tracking was rated marginally higher than eye tracking, with the exception of the most preferred font highlighting feature which may be reasoned by the aforementioned highlighting of a single word and, therefore, higher perceived accuracy of the eye tracker. Given the lack of clear preference between cursor- and eye-tracking interaction, these preliminary findings suggest that a low-cost reading support system without an eye tracker may offer a viable alternative (e.g., via a PDF reader or browser plugin). Being more accessible and cost-effective, cursor tracking could support wider adoption without compromising usability.

The key finding of the workshop was that the assistive reading system must be highly personalisable, enabling users with ADHD to adapt it to their individual needs. This finding aligns with the heterogeneous symptomatology of ADHD [29] and has also been highlighted as a critical consideration in assistive technology for people with autism, another neurodevelopmental disorder [30]. An initial step in this direction was the introduction of the line spacing adaptation option, which received positive feedback during the workshop, indicating the potential value of personalisation in digital reading support. Achieving an even higher level of personalisation could be attained by implementing Artificial Intelligence (AI) technologies. This would facilitate the realisation of automated adaptation mechanisms to users' reading performance. For instance, line jumps or the rereading of words could be identified during the reading process, allowing the system to adapt in real time by introducing visual reading aids (for more details, see [31]). When combined with other sensor modalities, such as infrared cameras capable of monitoring physical movement, machine learning methodologies can be used to detect sustained, increased physical activity while reading. Given evidence that movement changes with cognitive demand in people with ADHD [27][28], this approach could, for example, trigger reading support or provide feedback recommending a break when elevated movement suggests the task is becoming overly demanding. Further user-centred workshops and co-design sessions are needed to capture the experiences and priorities of users, supporting an iterative development process that ensures the resulting tools genuinely meet their needs.

The findings should be interpreted considering several limitations, including the small sample size, the inclusion of participants without formal ADHD diagnosis, and the exploratory nature of the workshop. In particular, it is important to bear in mind that our data is descriptive and exploratory. Thus, no statistical methods were applied. Additionally, the outcomes of the user workshop must be evaluated in conjunction with the current limitations of the prototype. One key factor is the accuracy of the eye tracker. A two-stage calibration was performed with all users to enable the most accurate tracking possible. If significant inaccuracies were detected after the

second stage, the process was repeated. Nonetheless, some inaccuracy could not be prevented, as the eye tracker exhibited substantial accuracy issues, particularly with two test subjects who wore their reading glasses over their regular glasses. This scenario must be avoided in future user tests. Furthermore, a deliberately implemented delay in all visual reading features, designed to accommodate rapid line jumps, was occasionally perceived as a system flaw rather than an intentional design choice. Hence, additional refinements to the visual design of the prototype and the incorporation of customisation options are imperative. Drawing on the findings, the following initial design principles for ADHD-specific reading assistance systems are proposed, each grounded in an identified theme and accompanied by a concrete design implication:

1) **Personalisation by Design**
   Reading assistance systems must provide adjustable parameters for visual features (e.g., colour, contrast, font, line spacing) and support the storage of individual user profiles to accommodate the heterogeneous needs and preferences of users with ADHD.
2) **Support Without Overload**
   Visual reading aids should be designed with restraint, as overly intense or persistent highlighting can increase cognitive load and distract from reading. Features such as blurring should be optional and configurable in intensity.
3) **Sustain Reading Flow Through Accurate Feedback**
   Features should provide continuous, low-latency feedback about the user's reading position (e.g., simultaneous word- and line-level highlighting) to maintain user trust and minimise disruption to reading flow.
4) **Ensure Accessible Interaction**
   Where possible, reading assistance systems should be usable on devices that users already own, without requiring the purchase of additional hardware or the attachment of equipment to the body. Cursor-based tracking offers a cost-effective and widely accessible entry point, while eye tracking should be considered an optional enhancement, for instance for attention-aware adaptation, rather than a prerequisite for system use.
5) **Enable Flexible Feature Combination**
   Users should be able to combine and layer features rather than selecting a single mode, enabling configurations that balance reading support with personal comfort and individual strategies.
6) **Integrate Adaptive Mechanisms Responsive to User State**
   Future iterations should consider integrating AI-based adaptation mechanisms, such as detecting line jumps, rereading behaviour, or elevated physical movement, to enable the system to provide contextually appropriate support dynamically during reading.

V. Conclusion and Future Work

The primary focus of the initial user-centred workshop was on user experience, as opposed to objective reading performance or long-term effects. Thus, the objective was to obtain an initial impression of usability, user experience, and acceptance. Additional research is needed to better understand the potential benefit of a reading support system for ADHD. Thus, as a next step, a randomised controlled trial will be conducted to compare an ADHD group with a neurotypical control group while reading digitally with the different visual support features. Multimodal (neuro-)physiological data, for instance brain, cardiac and body activity, will be collected during the study to investigate a potential benefit of visual reading support. Future work will explore adaptive extensions of the system, including context-aware mechanisms, such as motion tracking to dynamically trigger reading support or suggest breaks when cognitive demand increases. In addition, the integration of LLMs will be investigated for text simplification and summarising of challenging passages. Following the implementation of insights and feedback from the prototype testing, a subsequent user-centred workshop will be planned to further evaluate and refine the adjusted prototype. Overall, the initial assistive reading system developed for people with ADHD was perceived positively by the workshop participants. However, user workshop results have yielded clear opportunities for improvement, highlighting that while the prototype is functional in principle, additional refinement is required in order to better align with individual user needs and preferences.

ACKNOWLEDGEMENT

This work was supported by the Ministry of Culture and Science of the State of North Rhine-Westphalia as part of the project "Centre for Assistive Technology Rhine-Ruhr" (11/2023 to 10/2026, Grant No. PB22-076A). We thank all participants of the user workshop for their time and valuable feedback.

*KK and SB contributed equally to this work and share first authorship. All authors contributed to the manuscript. KK, SB, and HA designed and conducted the user-centred workshop; KK and HA analysed the data; AFK and SB developed the software. NW and CR supervised the work and coordinated the project.

References

[1] R. Molina et al., 'Children with attention-deficit/hyperactivity disorder show an altered eye movement pattern during reading', *Optom. Vis. Sci.*, vol. 97, no. 4, pp. 265–274, 2020. DOI: 10.1097/OPX.0000000000001498.

[2] P. Deans, L. O'Laughlin, B. Brubaker, N. Gay and D. Krug, 'Use of eye movement tracking in the differential diagnosis of attention deficit hyperactivity disorder (ADHD) and reading disability', *Psychology*, vol. 1, no. 4, pp. 238–246, 2010. DOI: 10.4236/psych.2010.14032.

[3] P. Stern, T. Kolodny, S. Tsafrir, G. Cohen and L. Shalev, 'Unique patterns of eye movements characterizing inattentive reading in ADHD', *J Atten Disord*, vol. 28, no. 6, pp. 1008–1016, Apr. 2024. DOI: 10.1177/10870547231223728.

[4] S. Caldani et al., 'Reading performance in children with ADHD: An eye-tracking study', *Ann. of Dyslexia*, vol. 72, no. 3, pp. 552–565, 2022. DOI: 10.1007/s11881-022-00269-x.

[5] C. Hanisch, R. Radach, K. Holtkamp, B. Herpertz-Dahlmann and K. Konrad, 'Oculomotor inhibition in children with and without attention-deficit hyperactivity disorder (ADHD)', *J Neural Transm*, vol. 113, no. 5, pp. 671–684, 2006. DOI: 10.1007/s00702-005-0344-y.

[6] P. B. Ulucan Atas, O. M. Ceylan, Y. E. Dönmez and O. Ozel Ozcan, 'Ocular findings in patients with attention deficit and hyperactivity', *Int Ophthalmol*, vol. 40, no. 11, pp. 3105–3113, 2020. DOI: 10.1007/s10792-020-01497-z.

[7] T. Banaschewski et al., 'Colour perception in ADHD', *Journal of Child Psychology and Psychiatry*, vol. 47, no. 6, pp. 568–572, 2006. DOI: 10.1111/j.1469-7610.2005.01540.x.

[8] A. Bellato et al., 'Association between ADHD and vision problems. a systematic review and meta-analysis', *Mol Psychiatry*, vol. 28, no. 1, pp. 410–422, 2023. DOI: 10.1038/s41380-022-01699-0.

[9] N. S. Bozhilova, G. Michelini, J. Kuntsi and P. Asherson, 'Mind wandering perspective on attention-deficit/hyperactivity disorder', *Neuroscience & Biobehavioral Reviews*, vol. 92, pp. 464–476, 2018. DOI: 10.1016/j.neubiorev.2018.07.010.

[10] P. Bonifacci, C. Viroli, C. Vassura, E. Colombini and L. Desideri, 'The relationship between mind wandering and reading comprehension: A meta-analysis', *Psychon Bull Rev*, vol. 30, no. 1, pp. 40–59, 2023. DOI: 10.3758/s13423-022-02141-w.

[11] E. Black and M. Hattingh, 'Assistive technology for ADHD: A systematic literature review', in *Innovative Technologies and Learning. ICITL 2020.*, ser. Lecture Notes in Computer Science, vol. 12555, Springer, Cham, 2020, pp. 514–523. DOI: 10.1007/978-3-030-63885-6_56.

[12] I. Svensson et al., 'Effects of assistive technology for students with reading and writing disabilities', *Disability and Rehabilitation: Assistive Technology*, vol. 16, no. 2, pp. 196–208, 2021. DOI: 10.1080/17483107.2019.1646821.

[13] D. R. Nuraini Herawati, W. Widajati and E. P. Sartinah, 'The role of text to speech assistive technology to improve reading ability in e-learning for ADHD students', *Journal of ICSAR*, vol. 6, no. 2, p. 169, 2022. DOI: 10 . 17977 / um005v6i22022p169.

[14] R. Tamdjidi and D. Pagès Billai, 'ChatGPT as an assistive technology to enhance reading comprehension for individuals with ADHD', retrieved: 2026.02.06, KTH Royal Institute of Technology EECS, School of Electrical Engineering and Computer Science, 2023. [Online]. Available: https://kth.diva-portal.org/smash/record.jsf?pid=diva2%3A1778288&dswid=-5755.html.

[15] H. Shen, O. Asiry, M. A. Babar and T. Bednarz, 'Evaluating the efficacy of using a novel gaze-based attentive user interface to extend ADHD children's attention span', *International Journal of Human-Computer Studies*, vol. 169, p. 102 927, 2023. DOI: 10.1016/j.ijhcs.2022.102927.

[16] D. Minas, E. Theodosiou, K. Roumpas and M. Xenos, 'Adaptive real-time translation assistance through eye-tracking', *AI*, vol. 6, no. 1, 2025. DOI: 10.3390/ai6010005.

[17] G. Schiavo, N. Mana, O. Mich, M. Zancanaro and R. Job, 'Attention-driven read-aloud technology increases reading comprehension in children with reading disabilities', *J. Comput. Assist. Learn.*, vol. 37, no. 3, pp. 875–886, 2021. DOI: 10.1111/jcal.12530.

[18] *Bionic reading*, retrieved: 2026.02.13, 2023. [Online]. Available: https://bionic-reading.com/.

[19] J. Snell, 'No, bionic reading does not work', *Acta Psychologica*, vol. 247, no. 7, p. 104 304, 2024. DOI: 10.1016/j.actpsy.2024.104304.

[20] J. K. B. Santos, 'Improving the reading comprehension of senior high school students through bionic reading', *Multidisciplinary International Journal of Research and Development*, vol. 4, no. 3, pp. 31–51, 2024, retrieved: 03.2026. [Online]. Available: https://www.mijrd.com/papers/v4/i3/MIJRDV4I30002.pdf.

[21] T. Jakovljević et al., 'The sensor hub for detecting the developmental characteristics in reading in children on a white vs. colored background/colored overlays', *Sensors*, vol. 21, no. 2, p. 406, 2021. DOI: https://doi.org/10.3390/s21020406.

[22] T. Jakovljević et al., 'The relation between physiological parameters and colour modifications in text background and overlay during reading in children with and without dyslexia', *Brain sciences*, vol. 11, no. 5, p. 539, 2021. DOI: 10.3390/brainsci11050539.

[23] R. C. Kessler et al., 'The world health organization adult adhd self-report scale (asrs): A short screening scale for use in the general population', *Psychological Medicine*, vol. 35, no. 2, pp. 245–256, 2005. DOI: 10.1017/s0033291704002892.

[24] I. Smythe, 'Adult checklist', 2001, retrieved: 03.2026. [Online]. Available: https://cdn.bdadyslexia.org.uk/uploads/documents/Dyslexia/Adult-Checklist-1.pdf?v=1554931003.

[25] D. L. Lefly and B. F. Pennington, 'Reliability and validity of the adult reading history questionnaire', *Journal of learning disabilities*, vol. 33, no. 3, pp. 286–296, 2000. DOI: 10.1177/002221940003300306.

[26] J. Brooke, 'Sus-a quick and dirty usability scale', *Usability evaluation in industry*, vol. 189, no. 194, pp. 4–7, 1996.

[27] M. D. Rapport et al., 'Hyperactivity in boys with attention-deficit/hyperactivity disorder (ADHD): A ubiquitous core symptom or manifestation of working memory deficits?', *J. Abnorm. Child Psychol.*, vol. 37, no. 4, pp. 521–534, 2009. DOI: 10.1007/s10802-008-9287-8.

[28] T. A. Hartanto, C. E. Krafft, A. M. Iosif and J. B. Schweitzer, 'A trial-by-trial analysis reveals more intense physical activity is associated with better cognitive control performance in attention-deficit/hyperactivity disorder', *Child Neuropsychol.*, vol. 22, no. 5, pp. 618–626, 2016. DOI: 10.1080/09297049.2015.1044511.

[29] Y. Luo, D. Weibman, J. M. Halperin and X. Li, 'A review of heterogeneity in attention deficit/hyperactivity disorder (ADHD)', *Front. Hum. Neurosci.*, vol. 13, p. 42, Feb. 2019.

[30] R. Cañete and M. E. Peralta, 'Applying technology to adapt assistive products to the sensorial characteristics of children with autism: A review', in *IEEE 11th International Conference on Serious Games and Applications for Health*, 2023, pp. 1–6. DOI: 10.1109/SeGAH57547.2023.10253810.

[31] A. F. Krause, K. Kannen, S. Büscher, C. Ressel and N. Wild-Wall, 'Pro-adaptive cognitive assistive technology: Concept and application in reading support for adhd', in *Extended Reality*, ser. Lecture Notes in Computer Science, L. T. D. Paolis, P. Arpaia and M. Sacco, Eds., vol. 15740, Springer, Cham, 2026, pp. 255–266. DOI: 10.1007/978-3-031-97772-5_17.

A Metacognitive Upstream Routing Framework for Accuracy Preservation and Computational Efficiency in Artificial Intelligence Systems

Naavya Shetty
Bachelor of Science in Computer Science and Philosophy
Department of Philosophy
University of Illinois Urbana-Champaign
Illinois, United States
e-mail: shetty.naavyasukesh@gmail.com

***Abstract*—Contemporary Artificial Intelligence (AI) systems often engage every input indiscriminately, resulting in unnecessary computation, unpredictable generalisation, and brittle behaviour on unfamiliar tasks. We present the Preprocessing Metacognitive System (PMS) 2.0, a system-agnostic metacognitive layer that evaluates incoming tasks and decides whether to accept, escalate, or refuse them before invoking any downstream reasoning system. PMS 2.0 seeks to provide interpretability at the level of computational governance - making transparent why the system chooses to engage, escalate, or refuse a task utilising confidence, task complexity, feasibility, novelty, and predicted benefit of escalation – to guide principled routing decisions without modifying downstream models. The system preserves conditional accuracy on escalated inputs, reduces computational load, and operationalises abstention as a first-class outcome, with previously refused tasks contributing to experience-informed efficiency gains. Evaluated across multiple domains and downstream architectures, PMS 2.0 demonstrates that metacognitive preprocessing can improve computational efficiency, reliability, and transparency, providing a practical framework for allocating resources where deliberative computation is most justified.**

Keywords-computational meta-reasoning; resource-rational AI; selective computation; input routing.

I. INTRODUCTION

Contemporary Artificial Intelligence (AI) systems remain remarkably indiscriminate in how they initiate and allocate computational effort. Once presented with an input, most models commit fully to processing it, regardless of whether the task is familiar, whether a lightweight heuristic could resolve it, or whether the system lacks the competence to address it reliably. This unfiltered mode of engagement produces several predictable failure modes: unnecessary computation on trivial problems, unpredictable generalisation on unfamiliar ones, and brittle or opaque behaviour when the system is operating outside its stable reasoning envelope. What is missing in nearly all modern systems is an architectural mechanism that evaluates incoming tasks before deeper computation begins - a preprocessing layer capable of deciding when to proceed, when to escalate, and when to refuse.

In an earlier work, we introduced the first version of such a mechanism: the Preprocessing Metacognitive System (PMS) 1.0 [1]. That preliminary proposal sketched the possibility of a metacognitive bottleneck that screens tasks, leverages previous experience, and restricts unnecessary reasoning. That paper established why deliberative pathways should not be treated as default channels, but it also left two questions open: whether this metacognitive bottleneck must be attached to a specific downstream model and how refusal, deferral, and computational justification could be operationalised beyond concept.

The present paper develops a significantly stronger formulation, PMS 2.0, which extends the original framework in several important ways, as seen in Table 1. First, PMS 2.0 formalises the bottleneck as a system-agnostic routing architecture capable of operating independently of downstream reasoning systems. Second, it introduces a small set of routing diagnostics, or interpretable meta-features, that characterise the relationship between the input and the system's known competence to guide routing decisions. Third, it operationalises computational abstention and refusal as explicit control actions, rather than treating them as failures. Finally, this paper provides the first empirical evaluation of the framework across multiple downstream architectures, demonstrating that metacognitive preprocessing can substantially reduce computational load while preserving conditional accuracy.

In this sense, PMS 2.0 is compatible with large language models, reinforcement-learning agents, symbolic planners, multimodal transformers, and classical machine-learning systems. Its role is not accuracy optimisation or predictive correction, but computational stewardship: deciding when deliberation is warranted, when fast acceptance is sufficient, and when a task must be refused on metacognitive grounds. It operates before these systems are activated, making judgments about the feasibility, novelty, and expected benefit. Simply put, we evaluate PMS 2.0 as a metacognitive control layer whose primary function is to decide when downstream computation is warranted. In this framework, abstention is not treated as an error, but as a deliberate control action.

The remainder of the paper proceeds as follows. Section 2 reviews the theoretical foundations and previous work on metacognition, dual-process reasoning, and resource-rational AI, situating PMS 2.0 within these publications. Section 3 introduces the architecture of PMS 2.0, detailing its modular components, metacognitive features, and system-agnostic routing principles. Section 4 presents the experimental design, including downstream tasks, evaluation metrics, and baseline comparisons. Section 5 reports results and discussion, demonstrating how PMS 2.0 preserves conditional accuracy, reduces computational load, operationalises refusals, and supports

 ISBN: 978-1-68558-375-0

TABLE I. COMPARISON OF PMS 1.0 AND PMS 2.0: KEY ARCHITECTURAL AND FUNCTIONAL EXTENSIONS

Feature	PMS 1.0	PMS 2.0
Metacognitive Bottleneck Concept	✓	✓
System-Agnostic Routing	×	✓
Explicit Meta-Features	×	✓
Computational Abstention	conceptual	operational
Empirical Evaluation	none	multi-model

transparent, interpretable decisions. Sections 6 and 7 outline limitations and avenues for future work, including extensions to learned controllers, improved novelty detection, and experience-informed refinement of refusal decisions. Section 8 concludes with broader implications for the implementation of metacognitive preprocessing in scalable AI systems.

II. LITERATURE REVIEW

The revised conception of PMS 2.0 emerges from three complementary research programs: dual-process theories of cognition, the science of metacognition and metareasoning, and formal models of bounded and resource-rational reasoning. Together, these frameworks provide the conceptual tools needed to reinterpret PMS 2.0 not as an intuitive idea, but as a theoretically necessary architectural component for intelligent and resource-limited agents.

A. Dual-Process Theory and Cognitive Control

Dual-process theory remains the most influential organising framework for thinking about the computational trade-offs between fast, pattern-driven cognition and slow, deliberative reasoning. Evans and Stanovich [2] characterise Type 1 processes as fast, automatic, and pattern-based, while Type 2 processes are slow, deliberative, and dependent on working memory. They emphasise that the Type 1 versus Type 2 distinction is most useful when treated as a difference in processing roles rather than as a set of anatomical modules. Debates about the empirical validity of dual-process distinctions often ask whether cognition is truly partitioned into two systems. De Neys [3] argues that dual-process distinctions remain indispensable in the functionalist sense even if the mechanisms themselves are interwoven. This shift toward functionality makes the dual-process vocabulary directly applicable to engineered systems, because designers can map computational primitives to the roles identified by the psychological literature.

B. Metacognition and Metareasoning

The work on metacognition and metareasoning provides the mechanisms by which an agent can inspect, evaluate, and regulate its own cognitive operations. Flavell's [4] early characterisation of metacognition as monitoring and control established the conceptual vocabulary, distinguishing knowledge about cognition from the processes that govern it. Building on that foundation, Cox [5] extended this to artificial systems, arguing that intelligent agents require explicit mechanisms to evaluate uncertainty, detect anomalies, and manage their computational investments. The edited volume by Cox and Raja [6] synthesises computational treatments of metacognition and metareasoning, translating philosophical and psychological concepts into concrete algorithmic concerns such as performance estimation, cost-aware control, and meta-level decision policies. More recently, efforts to formalise metareasoning ontologies have attempted to enumerate the primitive metacognitive functions and further refine the structure of metacognitive control by proposing a validated ontology of metareasoning operations, including confidence evaluation, cost estimation, and progress monitoring to implement self-aware systems [7]. These contributions are essential for designing layered control systems: they show both what needs to be tracked at the meta-level and how those signals can be operationalised.

C. Bounded Rationality and Resource-Rational Agents

Bounded rationality and heuristic decision-making provide the normative and descriptive rationale for preferring cheap, satisfying operations in many real-world environments. Simon's [8] original account framed decision making as a problem constrained by computational resources and environmental structure, introducing satisficing as a practical adaptive strategy. Gigerenzer and Todd [9] later demonstrated empirically how simple heuristics can outperform more complex calculations when resource costs and environmental regularities are considered; their work grounds the design choice to favour fast heuristics under realistic constraints. These strands converge in the resource-rational program, which formalises how an agent should allocate limited computation to maximise expected utility given processing costs.

D. Recent Work

There is a growing body of computational work that addresses how to learn or adapt metacognitive control. Schaeffer's [10] algorithmic treatments of metacognitive reinforcement learning show how a meta-level can learn escalation policies based on experience, defining metacognition as a learnable control problem rather than a set of fixed rules. Parallel research into practical metareasoning for modern models, such as recent demonstrations of value-of-computation ideas applied to large language models [11], indicates how cost-benefit reasoning can be scaled to modern and expensive systems.

Recent formalisations strengthen this foundation. Lieder and Griffiths [12] propose resource-rational analysis as a normative framework to optimize the use of computational resources. Russell and Wefald [13] extend this to artificial intelligence, offering formal principles for rational metareasoning that explicitly quantify accuracy-cost trade-offs. Rumana [14]

presents a synthesis of dual-process theory and metacognitive control, arguing that metacognitive mechanisms determine when Type-2 reasoning should intervene in Type-1 processing.

In the most recent computational turn, authors have attempted to map dual-process roles onto contemporary neuro-symbolic and machine learning systems. Gronchi and Perini [15] argue that dual-process distinctions naturally map onto neuro-symbolic architectures, where sub-symbolic networks provide intuitive judgments and symbolic systems support deliberation. Gronchi et al. [16] additionally show that inhibitory control mediates transitions between fast and slow processing in human cognition. These results suggest that engineering a gating mechanism is not only psychologically plausible but also architecturally sensible for hybrid AI systems.

On the tool side, scalable similarity search libraries like Facebook AI Similarity Search (FAISS) make experience-based novelty estimation tractable in large embedding spaces [17]. Taken together, these literatures provide a unified theoretical and practical toolkit for designing metacognitive preprocessors that are both principled and implementable.

III. Proposal

Building on the conceptual and empirical foundations summarised above, we propose PMS 2.0: a system-agnostic preprocessing metacognitive layer that makes principled, experience-informed routing decisions prior to invoking any downstream deliberative system. It also explicitly records refused inputs and periodically re-evaluates them after sufficient experience has accumulated, allowing the system to accept similar tasks more efficiently in the future.

The proposal integrates three core theoretical commitments: (1) a functional dual-process mapping that separates fast appraisal from deliberate inference, (2) explicit meta-level monitoring and cost-aware control, and (3) resource-rational decision rules that prioritise expected computational value.

A. Principles and goals

PMS 2.0 is designed around five guiding principles inherited from the literature. First, it implements a functional separation between rapid meta-appraisal and slow deliberation [2][3]. Second, it operationalises metacognition as monitoring plus control, tracking interpretable meta-features such as confidence, complexity, and feasibility rather than producing task-level predictions [4][5]. Third, PMS 2.0 evaluates the expected value of further computation in a resource-rational manner, comparing the expected improvement against cost [12][13]. Fourth, it uses experience-based similarity and novelty measures to detect unfamiliar inputs using scalable nearest-neighbour tools [17]. Fifth, the system emphasises transparency and decisional traceability: every routing decision is accompanied by meta-features and a human-readable reason, which supports interpretability and auditability [7].

B. Interpretable Meta-Features

A central design goal of PMS 2.0 is interpretability at the meta-decision level. In this work, interpretability refers to the property that the system's routing decisions can be directly explained in terms of human-understandable variables rather than opaque internal activations. The routing decisions in PMS 2.0 are governed by a small set of interpretable meta-features that describe the relationship between the incoming task and the system's accumulated experience. These features serve as meta-level diagnostics rather than task-level predictions. Their purpose is not to solve the task itself but to estimate whether further computation is justified.

- **Confidence** estimates the likelihood that a lightweight heuristic response would produce a reliable result. In practice, it reflects the consistency of the input with previously encountered patterns and the stability of heuristic outputs across similar cases. High confidence suggests that immediate acceptance or inexpensive processing may be sufficient.
- **Complexity** approximates the structural difficulty of the input relative to tasks the system has previously handled. This may reflect factors such as input length, compositional structure, or the number of interacting elements that must be processed. Higher complexity indicates that the task may require more deliberate reasoning or deeper computation.
- **Feasibility** represents the system's estimated ability to process the task successfully based on its existing competence. It is derived from similarity to previously solved cases and historical performance on related inputs. Low feasibility signals that the system may lack sufficient experience or capability to produce a reliable outcome.
- **Novelty** measures the distance between the current input and previously encountered tasks stored in the experience buffer. This signal is computed through similarity search in the embedding space and serves as an indicator of out-of-distribution inputs. High novelty encourages conservative behaviour, such as escalation or refusal.
- **Predicted Benefit of Escalation** ($\hat{\Delta}$) estimates the expected improvement in outcome quality that would result from invoking the downstream System-2 model. Escalation is therefore justified only when the anticipated benefit exceeds the expected computational cost.

Because these quantities are explicitly computed and logged, every routing decision produced by the MetaController can be traced back to a concrete set of meta-feature values. This transparency allows both developers and external auditors to inspect why a task was accepted, escalated, or refused without examining the internal structure of the downstream reasoning system. In this sense, interpretability in PMS 2.0 does not arise from explaining the reasoning of the downstream model itself, but from making the decision to invoke that reasoning transparent and inspectable.

C. Architectural Components

Concretely, PMS 2.0 consists of the following modular components, each directly motivated by the literature:

1) **Task encoder:** A modality-agnostic encoder produces compact embeddings that represent the structural features of the input. Representational choices follow the emphasis of

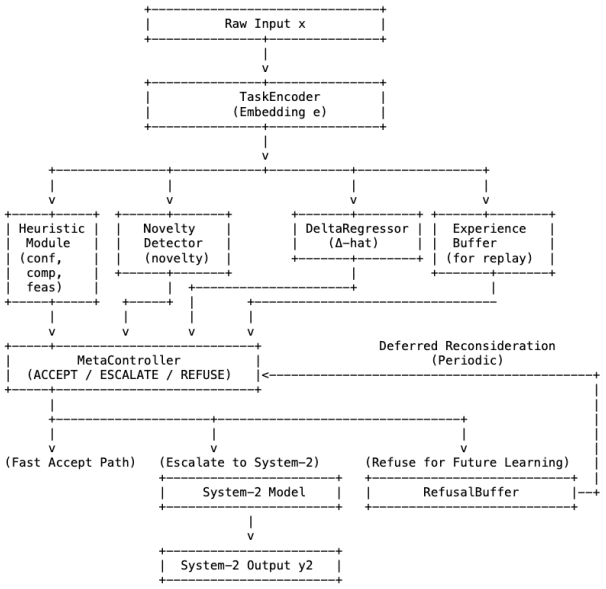


Figure 1. High-level Control Flow Representation

Russell and Norvig's [18] on appropriate feature spaces for downstream reasoning. The encoder normalises embeddings to permit consistent similarity computations and experience lookup.

2) **Heuristic meta-feature extractor:** A small, fast module produces interpretable meta-features, such as confidence, complexity, and feasibility, derived only from the embedding. This design follows Flavell's [4] monitoring/control distinction and is informed by Cox's [5] mapping of meta-level primitives to algorithmic constructs. The features are intentionally lightweight, so they can be produced cheaper than any downstream task computation.
3) **Novelty and experience module:** Using FAISS or equivalent nearest-neighbour search, the system evaluates the distance to stored experiences to yield a novelty score. This score operationalises epistemic unfamiliarity and provides a principled signal for conservative behaviour when the system encounters out-of-distribution inputs [17].
4) **Delta or value regressor:** A learned regressor estimates the expected benefit ($\hat{\Delta}$) of invoking the downstream system. This module formalises the resource-rational imperative described by Lieder and Griffiths [12] and Russell and Wefald [13]: escalation is justified only when the expected gain exceeds the anticipated cost of computation. Training uses logged escalations to supervise $\hat{\Delta}$ learning, in effect implementing a form of metacognitive reinforcement learning [10].
5) **MetaController (decision policy):** For transparency and robustness, the first instantiation of the controller is rule-based:
   - ACCEPT when $\hat{\Delta} \leq 0$ and feasibility is high
   - ESCALATE when $\hat{\Delta}$ > threshold or novelty is high
   - REFUSE when feasibility or confidence are extremely low
   - Refused inputs are stored in a RefusalBuffer, which preserves embeddings, meta-features, and the original input. Periodically, these entries are reconsidered: novelty and $\hat{\Delta}$ are recalculated, and the MetaController may update previous REFUSE decisions. This mechanism allows PMS 2.0 to gradually accept inputs that were initially deferred, improving overall computational efficiency while maintaining system-agnostic metacognitive control. The REFUSE action enables the most distinctive function of PMS 2.0: the ability to say no to inputs when the system cannot guarantee reliable processing, while also influencing further learning. This choice relies on the interpretability emphasised in metareasoning ontologies while leaving open the option of replacing the rule-based controller with a learned policy once enough experience is available [7][10].

The overall routing architecture of PMS 2.0 is illustrated in Figure 1, which shows how incoming inputs are encoded, evaluated using meta-features and novelty estimation, and subsequently routed by the preprocessing router MetaController to either ACCEPT, ESCALATE, or REFUSE pathways.

D. Design justification and theoretical fit

This architecture is the natural engineering embodiment of the earlier surveyed literature. The functional dual-process mapping prescribes a fast appraisal stage; metareasoning and metacognition provide the primitives needed to produce that appraisal; bounded rationality and resource-rational analysis provide the decision criterion that converts appraisal into action. Experience-based novelty detection operationalises cautious behaviour in the presence of distribution shift, a practice recommended by contemporary neuro-symbolic mappings and empirical meta-analyses [16][15]. Importantly, this design choice intentionally separates interpretability from task-level prediction, and the system is deliberately agnostic about the identity or internals of System-2. The adapter interface allows any downstream model to be invoked as a black box, which not only respects constraints common in production systems where internals are proprietary or too large to modify, but also instead provides an interpretable control layer that governs when such models should be invoked. By grounding routing decisions in explicit meta-features, the system produces auditable decision traces that support debugging, accountability, and principled resource allocation.

E. Operational behaviour and evaluation

Practically, PMS 2.0 functions as an input router. For each input, the encoder and meta-features are computed; novelty is assessed with respect to an experience buffer; $\hat{\Delta}$ is predicted; the MetaController issues ACCEPT, ESCALATE, or REFUSE; if ESCALATE is chosen, the system calls System-2 and records the result, feeding it back to the experience buffer and using it to refine $\hat{\Delta}$. Periodically, the RefusalBuffer is revisited: novelty and $\hat{\Delta}$ are recalculated for refused inputs, and the MetaController may revise prior REFUSE decisions. The

success criteria are twofold: measurable reductions in expensive System-2 calls (computational savings) and preservation of correctness on escalated cases (no unjustified degradation). Efficiency gains are enhanced because previously refused inputs can later be processed using learned experience, reducing repeated System-2 calls. Secondary criteria include improved transparency (decision logs) and meaningful refusal behaviour (declines those that indicate high novelty or low feasibility).

F. Practical considerations

Several design choices are directly derived from the literature and practical constraints. First, meta-features should be interpretable and cheap, so they do not erase the savings introduced by avoiding System-2 calls. Second, the DeltaRegressor must be trained from a representative set of escalations; this creates a bootstrap period during which the controller relies more heavily on conservative thresholds. Third, because novelty metrics based on embedding distances can be brittle in some multimodal or highly structured domains, future work should investigate learned density estimators or contrastive representations as alternatives to better detection of semantic novelty [17][19].

IV. EXPERIMENTATION

This section describes the experimental setup used to evaluate PMS 2.0, including implementation details, evaluation protocols, and baseline comparisons.

A. Implementation Details

The PMS 2.0 is implemented in PyTorch with modular subcomponents designed for easy replacement or extension. The design principles guiding the implementation reflect the architectural commitments described above: system-agnosticism, transparency, and strict separation of meta vs task processing.

Figure 2 shows the modular software organisation of PMS 2.0, which outlines the directory structure used to separate the metacognitive modules, System-2 adapter, training utilities, and experience buffers.

```
pms/
├── models/
│   ├── task_encoder.py
│   ├── heuristic_module.py
│   ├── delta_regressor.py
│   ├── meta_controller.py
│   └── system2_adapter.py
├── core/
│   └── experience_buffer.py
│   └── refusal_buffer.py
├── utils/
│   └── novelty_detector.py
├── router/
│   └── router.py
└── training/
    └── supervised_trainer.py
```

Figure 2. Project Directory Structure

- Each module corresponds to a theoretical construct. For example, the HeuristicModule implements the fast metacognitive appraisal system and outputs only meta-features; the System2Adapter wraps any arbitrary model through a uniform interface.
- **Experience Buffer and FAISS Integration:** The ExperienceBuffer leverages FAISS for large-scale similarity search, allowing efficient novelty estimation and replay sampling. Each stored experience contains:
  - embedding
  - confidence, complexity, feasibility
  - novelty
  - System-2 output (if escalated)
  - $\hat{\Delta}$ (actual benefit when calculable)

  This reflects the metacognitive RL framework in Schaeffer [10], but, more importantly, PMS 2.0 does not require task labels. The system learns entirely from meta-level outcomes, not predictions, further distancing it from the supervised metacognition design in my previous work [1].
- **System2 Abstraction Layer**: The System2Adapter ensures that PMS 2.0 does not assume anything about the underlying reasoning system. Any callable model that accepts an input tensor and returns an output can be integrated. This keeps the metacognitive layer decoupled and portable.
- **Training Infrastructure:** The supervised trainer prepares the data for training the DeltaRegressor. The novelty, confidence, complexity, feasibility, and embeddings are all fed as features; $\hat{\Delta}$ is used as the target when available. The system can optionally use the accumulation of online experiences, expanding its applicability to continuous learning settings.

B. Experimental Setup

To evaluate PMS 2.0, two complementary goals guide the experimental design:

1) Does PMS 2.0 improve system reliability, transparency, and the ability to reject malformed inputs?
2) Does PMS 2.0 reduce computational load by diverting low-value cases away from System-2 models?

We tested PMS 2.0 as a preprocessing layer in front of an arbitrary System-2 model. In line with its system-agnostic design, the experiments do not rely on a particular choice of System-2. Instead, we evaluated PMS 2.0 on different downstream architectures. Evaluation also tracks the system's ability to revisit previously refused inputs, measuring acceptance rate and computational savings on formerly deferred tasks.

C. Systems Under Evaluation

We evaluate PMS 2.0 paired with:

- A medium-scale transformer model (for text inputs)
- A ResNet classifier (for image inputs)
- A symbolic reasoning engine (for structured tasks)

This diversity supports the claim that PMS 2.0 is task-independent and plug-compatible with modern AI systems.

D. Data and Task Conditions

For each System-2 model, the inputs are divided into:

- Routine / in-distribution cases
- Edge-case high-complexity tasks

- Adversarily perturbed or malformed inputs
- Completely novel inputs (unseen embedding clusters)

These categories allow for the evaluation of PMS 2.0's routing decisions: can it meaningfully refuse, accept, or escalate tasks.

E. Training the Delta Regressor

During preliminary runs, System-2 is queried on a subset of inputs to compute the actual deltas to train the DeltaRegressor. The rest of the PMS 2.0 components are frozen. This reflects the metacognitive RL regime in Schaeffer [10], but without requiring action-level reward functions.

F. Evaluation Metrics

We evaluate the following metrics:

1) Escalation Rate – How often inputs are forwarded to System-2.
2) Refusal Accuracy – Ability to correctly reject unprocessable or harmful inputs.
3) Coverage Preservation – Fraction of correct System-2 outputs preserved when PMS 2.0 is inserted.
4) Computational Savings – Reduction in System-2 calls.
5) Transparency Metrics – Distributions of confidence, novelty, feasibility, and $\hat{\Delta}$ for each routing class.
6) Error Mitigation – Reduction in System-2 hallucinations or unsafe outputs relative to baseline.
7) Deferred Acceptance Rate – fraction of previously refused inputs that are eventually escalated and accepted successfully.

G. Baselines

We compare against:

- System-2 alone (no PMS 2.0).
- Simple threshold-based gating system.
- Heuristic-only version of PMS 2.0 (no $\hat{\Delta}$ model).

This ensures that the improvements are attributable to the metacognitive layer rather than trivial heuristics.

V. Results | Discussion

The evaluation examined how well PMS 2.0 functions as a system-agnostic upstream routing layer that governs when a more expensive or more capable downstream AI system (System-2) should be invoked. PMS 2.0 was assessed on four primary dimensions: routing efficiency, conditional and unconditional accuracy, error exposure, and transparency of decisions. In this evaluation, computational cost is approximated by the number of invocations of the downstream System-2 model, which represents the dominant source of computational expense in most AI pipelines. Comparisons were conducted against three baselines: unconditional System-2 invocation, a simple threshold-based gating system, and a heuristic-only variant of PMS without learned benefit estimation.

Experiments were conducted on 200 held-out synthetic inputs using a lightweight and moderately error-prone System-2 model. This controlled setting was intentionally selected to isolate the effects of metacognitive routing from downstream model capacity. All systems were evaluated under identical conditions, with PMS 2.0 operating strictly upstream and without access to the System-2 internals.

Across the evaluation set, as shown in Table 2, PMS 2.0 reduced System-2 calls from 120 to 17, corresponding to an 85.8% reduction in downstream computation relative to unconditional escalation. When PMS 2.0 authorised escalation, conditional accuracy, defined as correctness on escalated inputs only, remained comparable to the baseline systems. By construction, PMS 2.0 introduces a distinction between conditional accuracy and unconditional accuracy, where unconditional accuracy is measured over all inputs, with refusals conservatively counted as incorrect. As coverage decreases, unconditional accuracy correspondingly declines, reflecting intentional abstention rather than degraded prediction quality.

This distinction is central in interpreting the behaviour of PMS 2.0. The system is not designed to maximise unconditional accuracy, as doing so would require escalating on every input. Instead, PMS 2.0 explicitly regulates when deliberative computation is warranted. Inputs that are not escalated are refused rather than guessed, making abstention a first-class outcome rather than an implicit failure. In safety-critical or resource-constrained settings, such behaviour can be preferable to unexamined engagement.

From a computational perspective, this reduction is significant because System-2 invocation represents the most expensive stage of the processing pipeline. By filtering inputs upstream, PMS 2.0 limits expensive reasoning to a small subset of cases where additional computation is predicted to be beneficial. In the present evaluation, this results in over six-fold fewer System-2 calls while maintaining comparable conditional accuracy on the escalated tasks. These results provide empirical support for the central hypothesis of the paper: that metacognitive preprocessing can substantially reduce computational expenditure without degrading performance on tasks that genuinely require deeper reasoning.

Efficiency metrics further highlight the benefits of metacognitive routing. Although unconditional System-2 invocation and threshold-based gating achieve full coverage, they incur maximal downstream cost because every input is forwarded to the expensive reasoning system and, in the case of threshold gating, substantially increase error exposure. In contrast, PMS 2.0 significantly reduces downstream computation while also lowering the absolute number of downstream errors by selectively refusing inputs likely to result in misclassification. As a result, PMS 2.0 not only reduces computation but also actively limits the propagation of downstream failures by preventing the System-2 model from engaging with inputs outside its competence.

The heuristic-only PMS variant achieved the highest conditional accuracy among all systems, suggesting that static heuristics can be effective in constrained settings. However, this variant lacks adaptive benefit estimation and exhibits a higher error exposure than the full PMS 2.0 system. This result highlights a key trade-off: learned benefit estimation prioritises conservative escalation and cost reduction, sometimes at the expense of coverage. PMS 2.0 makes this trade-off explicit

TABLE II. EVALUATION RESULTS: COMPARISON OF ROUTING STRATEGIES ACROSS ACCURACY, COVERAGE, AND SYSTEM-2 UTILISATION

System	Cond. Acc.	Uncond. Acc.	Coverage	S2 Calls	Savings	Acc./S2 Call
System-2 Only	0.0750	0.0750	100.0%	120	0.0%	0.0750
Threshold Gate	0.0750	0.0750	100.0%	120	0.0%	0.0750
Heuristic-Only PMS	**0.0886**	0.0583	65.8%	79	34.2%	**0.0886**
Full PMS 2.0	0.0588	0.0083	14.2%	17	**85.8%**	0.0588

and measurable.

In addition to routing decisions, PMS 2.0 produced fully transparent decision logs. Each ACCEPT, ESCALATE, or REFUSE decision was accompanied by the underlying values of metacognitive characteristics and a human-readable explanation. This transparency was achieved without modifying or inspecting System-2, supporting the claim that PMS 2.0 can function as an external accountability layer for black-box models.

A key lesson from these experiments is that aggressive computational savings can substantially reduce coverage if the benefit estimate is conservative. Although PMS 2.0 achieved the greatest reduction in System-2 usage and downstream error exposure, it underperformed the heuristic-only variant in conditional accuracy. This failure mode motivates future work on adaptive threshold calibration, online learning, and multi-tier routing architectures. Overall, the results confirm that PMS 2.0 functions as intended: reducing unnecessary computation while preserving correctness on authorised escalations, and operationalising abstention as a principled metacognitive decision rather than a failure of prediction.

VI. LIMITATIONS

Although PMS 2.0 demonstrates significant computational savings and principled abstention behaviour, several limitations constrain the scope of the current results.

- **Synthetic evaluation setting:** The evaluation was performed on synthetic data using a lightweight synthetic System-2 model. This choice enabled controlled experimentation and clear attribution of effects, but it limits the ecological validity of the results. Real-world deployments will involve substantially more complex tasks, heterogeneous input distributions, and significantly more expensive downstream models. Although routing principles are expected to generalise, empirical trade-offs between coverage, accuracy, and efficiency may differ across domains.
- **Task-specific benefit definition:** PMS 2.0 relies on a user-defined benefit function to train the DeltaRegressor. This design provides flexibility, but it also restricts its applicability to settings where downstream quality can be explicitly quantified. In domains where performance objectives are ambiguous or multi-dimensional, benefit estimation may be difficult to specify reliably, increasing sensitivity to miscalibration and conservative routing behaviour.
- **Rule-based MetaController:** The current MetaController uses fixed, manually selected thresholds to determine ACCEPT, ESCALATE, and REFUSE actions. Although this choice improves transparency and interpretability, it limits adaptability across tasks and environments. As observed in the experiments, conservative thresholding can lead to excessive refusals and reduced conditional accuracy, particularly when benefit estimates are uncertain.
- **Simplified novelty estimation:** Novelty detection is implemented using distance-based similarity in the embedding space. This approach is effective for small experience buffers, but it may inadequately capture semantic or functional novelty in highly structured, multimodal, or high-dimensional input spaces. As a result, novelty signals may be coarse or unreliable in more complex deployments.
- **Cold-start effects and limited early experience:** PMS 2.0 requires an experience buffer populated through interaction with System-2 to train the DeltaRegressor. During early operation, benefit estimates are, therefore, based on sparse data, which can amplify conservative routing decisions. Although the system logs refused inputs in a RefusalBuffer for deferred reconsideration and gradual incorporation into experience, early refusals may still be weakly informed until sufficient interaction data accumulate.

VII. FUTURE WORK

The experimental results highlight several concrete directions for improving PMS 2.0, particularly in response to the observed trade-off between aggressive computational savings and conditional accuracy.

1) **Adaptive threshold calibration and learned control:** The experiments show that the current rule-based MetaController produces strong computational savings but may be overly conservative in escalation decisions. This motivates replacing the fixed rule-based MetaController with an adaptive or learned controller. A policy trained via offline decision modelling or reinforcement learning could dynamically adjust escalation and refusal thresholds to balance efficiency against conditional accuracy, while retaining interpretability through explicit explanation mechanisms.
2) **Refined benefit estimation:** The system's reliance on a task-specific benefit function provided by the user enables flexibility, but it also increases sensitivity to conservative or miscalibrated benefit estimates, as reflected in reduced coverage. Future work should explore domain-agnostic or learned surrogate objectives that better align predicted benefit with downstream accuracy gains, reducing over-refusal without reverting to indiscriminate escalation.
3) **Multi-tier routing architectures:** The sharp coverage drop observed under aggressive routing suggests that binary escalation decisions may be overly coarse for many real world environments. Extending PMS 2.0 to support multi-level routing across several downstream models with graded cost

and capability profiles could mitigate this effect, allowing conservative refusals to be replaced with intermediate, lower-cost escalation options.

4) **Online learning and refusal reuse:** PMS 2.0 currently trains its benefit estimator offline and logs refusals without fully exploiting them during the evaluation. A fully online variant could continuously update metacognitive estimates, revisit previously refused inputs, and incorporate explicit refusal regret signals. This would allow conservative refusals to be corrected over time, improving conditional accuracy while preserving computational savings, and enable the system to progressively expand its competence over time.

These directions aim to transform PMS 2.0 from a static routing mechanism into a continuously adapting metacognitive control layer capable of allocating computational resources efficiently across a wide range of AI systems.

VIII. Conclusion

The Preprocessing Metacognitive System presented in this work provides a theoretically grounded and implementable blueprint for a metacognitive preprocessing layer. It synthesises insights from dual-process theory, computational metareasoning, and resource-rational analysis into a modular architecture designed for transparency and practical integration. In doing so, it operationalises the normative insight that computation itself is a scarce resource and that intelligent systems should explicitly decide when additional computation is justified.

By placing metacognitive evaluation at the preprocessing stage, PMS 2.0 enables intelligent routing without inspecting or modifying downstream models. The system is fully system-agnostic, interpretable, and capable of principled abstention. The empirical results demonstrate that this approach can substantially reduce downstream computational cost while preserving correctness on authorised escalations and providing complete auditable decision explanations.

Importantly, PMS 2.0 treats rejections as provisional rather than terminal. Declined inputs are logged and may be reconsidered as additional experience accumulates, allowing previously deferred cases to be escalated efficiently when predicted benefit increases. This deferred handling reframes abstention as a dynamic, experience-informed mechanism rather than a static rejection policy, supporting gradual improvement in routing decisions over time without redundant computation.

More broadly, the architecture illustrates how metacognition can function as a practical control layer for modern AI systems. Rather than relying exclusively on larger models or increased data, PMS 2.0 demonstrates that explicitly reasoning about when to compute can meaningfully improve system efficiency and reliability. Although the current evaluation is limited in scope, the results suggest that metacognitive control at the preprocessing-level offers a promising and extensible approach to managing cost, risk, and accountability in scalable AI pipelines.

References

[1] N. Shetty, "Metacognition-driven preprocessing for optimized artificial intelligence performance", in *Proceedings of the Seventeenth International Conference on Advanced Cognitive Technologies and Applications (COGNITIVE 2025)*, IARIA Press, 2025, ISBN: 978-1-68558-260-9. [Online]. Available: https://www.thinkmind.org/library/COGNITIVE/COGNITIVE_2025

[2] J. S. B. T. Evans and K. E. Stanovich, "Dual-process theories of higher cognition: Advancing the debate", *Perspectives on Psychological Science*, vol. 8, no. 3, pp. 223–241, 2013. DOI: 10.1177/1745691612460685

[3] W. De Neys, "Dual-process theory 2.0", *Routledge*, 2017, Edited volume.

[4] J. H. Flavell, "Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry", *American Psychologist*, vol. 34, no. 10, pp. 906–911, 1979. DOI: 10.1037/0003-066X.34.10.906

[5] M. T. Cox, "Metacognition in computation: A selected research review", *Artificial Intelligence*, vol. 169, no. 2, pp. 104–141, 2005. DOI: 10.1016/j.artint.2005.10.004

[6] M. T. Cox and A. Raja, Eds., *Metareasoning: Thinking about thinking*. Cambridge, MA: MIT Press, 2011. DOI: 10.7551/mitpress/9780262014809.001.0001

[7] M. F. Caro, M. T. Cox, and R. E. Toscano-Miranda, "A validated ontology for metareasoning in intelligent systems", *Journal of Intelligence*, vol. 10, no. 4, p. 113, 2022. DOI: 10.3390/jintelligence10040113

[8] H. A. Simon, "Rational choice and the structure of the environment", *Psychological Review*, vol. 63, no. 2, pp. 129–138, 1956.

[9] P. M. Todd, G. Gigerenzer, and the ABC Research Group, *Simple Heuristics That Make Us Smart*. Oxford University Proceedings, Jan. 1999.

[10] R. Schaeffer, "An algorithmic theory of metacognition in minds and machines", arXiv:2111.03745, 2021. [Online]. Available: https://arxiv.org/abs/2111.03745

[11] C. N. De Sabbata, T. R. Sumers, B. AlKhamissi, A. Bosselut, and T. L. Griffiths, "Rational metareasoning for large language models", arXiv:2410.05563, 2024. [Online]. Available: https://arxiv.org/abs/2410.05563

[12] F. Lieder and T. L. Griffiths, *Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources*. Behavioral and Brain Sciences, 2018.

[13] S. J. Russell and E. Wefald, "Principles of metareasoning", *Artificial Intelligence*, vol. 49, no. 1–3, pp. 361–395, 1991. DOI: 10.1016/0004-3702(91)90009-U

[14] A. Rumana, "Metacognitive control in single- vs. dual-process theory", *Thinking and Reasoning*, vol. 29, no. 2, pp. 177–212, 2023. DOI: 10.1080/13546783.2022.2047106

[15] G. Gronchi and A. Perini, "Dual-process theories of thought as potential architectures for developing neuro-symbolic ai models", *Frontiers in Cognition*, vol. 3, pp. 1–5, 2024. DOI: 10.3389/fcogn.2024.1356941

[16] G. Gronchi, G. Gavazzi, M. P. Viggiano, and F. Giovannelli, "Dual-process theory of thought and inhibitory control: An ale meta-analysis", *Brain Sciences*, vol. 14, no. 1, p. 101, 2024. DOI: 10.3390/brainsci14010101

[17] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with gpus", *IEEE Transactions on Big Data*, vol. 7, no. 3, pp. 535–547, 2019. DOI: 10.1109/TBDATA.2019.2921572

[18] S. J. Russell and P. Norvig, *Artificial intelligence: A modern approach*, 4th ed. London: Pearson, 2021.

[19] Sun, "Contents", in *The Cambridge Handbook of Computational Cognitive Sciences* (Cambridge Handbooks in Psychology), Cambridge Handbooks in Psychology. Cambridge University Press, 2023, pp. v–viii.

Comparative Evaluation of RAG and GraphRAG for Open-Ended Question Answering

Jadesola Osinowo
SCEIS
Ulster University
Londonderry, United Kingdom
e-mail: j.osinowo@ulster.ac.uk

Abiodun Adebayo
SCEIS
Ulster University
Londonderry, United Kingdom
e-mail: a.adebayo@ulster.ac.uk

Sonya Coleman
SCEIS
Ulster University
Londonderry, United Kingdom
e-mail: sa.coleman@ulster.ac.uk

Dermot Kerr
SCEIS
Ulster University
Londonderry, United Kingdom
e-mail: d.kerr@ulster.ac.uk

Justin Quinn
SCEIS
Ulster University
Londonderry, United Kingdom
e-mail: jp.quinn@ulster.ac.uk

***Abstract*—Retrieval-Augmented Generation (RAG) has become the basis in modern question answering (QA) systems, combining Large Language Models (LLMs) with external document retrieval. However, traditional RAG architectures often struggle with retrieving semantically structured or context-rich content, which can impact accuracy and relevance. This paper presents a comparative evaluation of a standard RAG model and a GraphRAG model. The GraphRAG approach combines graph-based document representation within the retrieval pipeline to assess their efficiency on a structured question answering task. This research leverages two modern Ollama-hosted models, Phi-4 and LLaMA 3.2 3B, as the base language models for both retrieval pipelines. Using a custom dataset derived from German coalition policy documents, this research evaluates performance through both lexical and semantic metrics. The results demonstrate that GraphRAG consistently outperforms traditional RAG in semantic alignment and contextual accuracy, particularly when paired with Phi-4. These findings aim to contribute to the growing body of work on hybrid retrieval strategies and support the case for graph-enhanced architectures in long-form QA systems, which are central to advancing structured and knowledge-aware retrieval methods in complex information domains.**

Keywords- RAG; GraphRAG; LLM; GenAI.

I. INTRODUCTION

Retrieval-Augmented Generation (RAG) has emerged as a foundational method to augment large language models (LLMs) with factual grounding, enabling them to access and condition on external knowledge sources at inference time. By retrieving relevant documents and guiding the generation process with this evidence, RAG bridges the strengths of information retrieval systems and generative language models, reducing hallucinations and improving factual accuracy [1]. However, traditional RAG systems typically treat retrieved text as flat sequences, overlooking deeper semantic, structural, or logical relationships within the context. This limitation has been increasingly recognised as a barrier in handling complex or multi-hop queries, especially across dense knowledge domains like legal, policy, or scientific corporations [2].

To address this, recent innovations such as GraphRAG propose a more structured form of context representation. Rather than presenting documents as flat chunks, GraphRAG represents retrieved information as a graph of entities, concepts, and relationships, enhancing reasoning over inter-related content [3]. This graph-based approach has demonstrated empirical benefits in use cases such as Biomedical Question Answering (BQA), complex multi-hop QA, and long context reasoning tasks [4][13] outperforming traditional RAG in both factual alignment and coherence [4][5]. Despite these promising developments, comparative evaluations remain scarce particularly across different LLM backbones and real-world structured datasets.

For example, in a policy-related query requiring evidence from both cybersecurity and economic governance sections, a traditional RAG system retrieves the top-k most similar text chunks independently. In contrast, GraphRAG traverses relational links between entities across documents, enabling multi-hop integration of evidence before generation. This structured traversal allows the model to construct a more coherent and contextually connected response.

This paper evaluates the performance of traditional RAG and GraphRAG architectures across two modern open-source LLMs hosted via Ollama: Phi-4, a compact model optimised for high-context reasoning, and LLaMA 3.2 3B, known for its larger parameter count and general knowledge coverage. Using a custom benchmark of policy-related QA tasks derived from coalition agreement documents, we assess how well each model-architecture pair performs in grounded, retrieval-heavy generation scenarios. To achieve a holistic evaluation, we employ both lexical overlap metrics (BLEU, ROUGE) and semantic alignment scores using RAGAS including metrics like cosine similarity, faithfulness, and context relevance. The results show that GraphRAG consistently outperforms traditional RAG pipelines, particularly when paired with Phi-4, generating answers that are both more faithful to the source materials and better aligned with user queries. The remainder of this paper is structured as follows: Section II describes the methodology and system architecture, Section III outlines the experimental setup and evaluation framework, Section IV presents the performance results and analysis, and Section V concludes with key findings and future research directions.

II. RELATED WORK

Retrieval-Augmented Generation (RAG) has been extensively studied to integrate information retrieval with large language models, enabling systems to leverage external knowledge at inference time. Previous works have demonstrated the potential for RAG in question answering (QA), summarisation, and dialogue tasks, improving factual grounding compared with standalone LLMs [1]. However, limitations such as shallow retrieval granularity and lack of semantic structuring often reduce performance when queries require multi-hop reasoning or relational context [7].

To overcome these limitations, graph-based retrieval methods have emerged. GraphRAG extends a traditional RAG by organising documents into graph structures, where nodes represent entities or semantic units and edges encode relationships such as co-occurrence, hierarchy, or topical linkage [8]. This approach allows retrieval to follow relational paths, yielding context that is both more compact and semantically meaningful. Recent comparative studies highlight GraphRAG's advantage in tasks requiring multi-document reasoning, such as query-based summarisation or knowledge graph QA, while traditional RAG remains competitive on simpler single-hop factoid tasks [9][10].

A. Traditional RAG Structure

Traditional RAG systems follow a relatively linear pipeline, as illustrated in Figure 1:

1) *Document Preprocessing and Chunking:* Source documents are segmented into overlapping text chunks, often based on fixed token or character lengths [1].
2) *Indexing and Retrieval:* A retriever (sparse, dense, or hybrid) determines the top-*k* chunks most relevant to a query. Dense retrieval methods such as sentence embeddings have become standard, improving semantic recall [11].
3) *Augmented Generation:* Retrieved chunks are concatenated and passed to the LLM to generate an answer.

This structure has been validated across multiple domains, but its reliance on flat sequences introduces redundancy and noise. When chunks lack explicit relational encoding, models may overfit on superficial overlaps (e.g., lexical similarity) rather than extracting deeper conceptual links [7]. Furthermore, traditional RAG pipelines often perform poorly in domains with long, interdependent texts, where hierarchical relationships between concepts are essential.

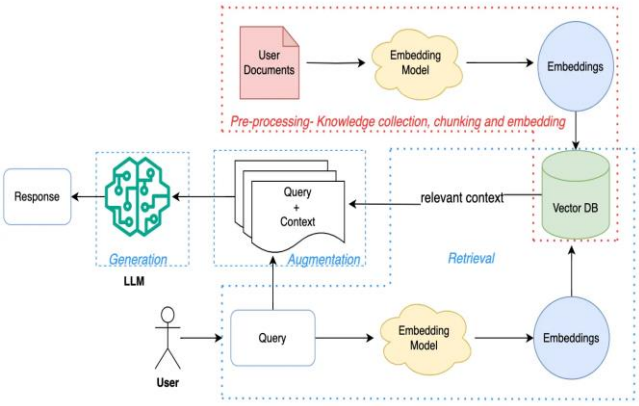


Figure 1. Illustration of RAG Architecture [30].

B. GraphRAG Structure

GraphRAG modifies the retrieval stage by introducing graph-based indexing and context construction. Instead of treating documents as isolated chunks, GraphRAG builds graphs where:

- Nodes correspond to entities, sentences, or topical communities.
- Edges capture relationships such as semantic similarity, co-reference, or hierarchy.
- Community Reports or hierarchical summaries can be generated for high-level nodes to reduce redundancy [8].

At inference time, queries are mapped onto a graph, and retrieval proceeds by exploring neighbourhoods or hierarchical paths. As illustrated in Fig. 2, query encoding, graph traversal, and subgraph selection determine the context passed to the language model, enabling structurally coherent context construction prior to generation. This results in evidence that is both structurally coherent and semantically accurate. For example, in multi-hop question answering, GraphRAG can traverse connections between entities across documents, yielding more precise retrieval contexts than top-k flat retrieval [9].

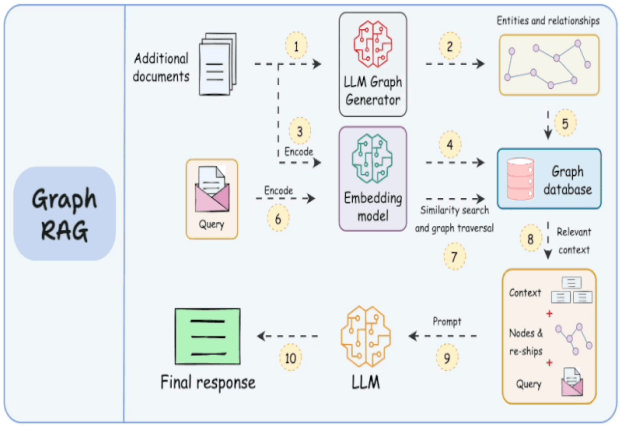


Figure 2. Illustration of GraphRAG Architecture [31].

C. Trend of Traditional RAG Research

Since its inception, Retrieval-Augmented Generation (RAG) has evolved from a hybrid of retrieval systems and neural generation into a widely adopted architecture for knowledge intensive tasks [1]. Foundational work formalised the integration of dense retrieval with sequence generation, addressing the limitations of parametric memory in LLMs by enabling access to external corpora at inference time [1][2]. The classical RAG pipeline consists of a dense retriever (e.g., Dense Passage Retrieval or vector search) that embeds queries and document chunks into a shared semantic space, followed by a generative model that conditions on the retrieved context to produce grounded outputs [10][11].

Subsequent research refined this basic model in two main ways. First, retrieval architectures were improved through hybrid sparse–dense retrievers and reranking mechanisms to enhance precision and recall, particularly in long-tail or specialised domains [10][16]. These approaches balance lexical matching with semantic similarity. Second, adaptive and feedback-driven retrieval methods emerged, where initial retrieval results guide query reformulation or iterative retrieval loops, supporting multi-step reasoning in complex QA tasks [16][17].

In parallel, work on generator–retriever integration introduced fusion mechanisms beyond simple concatenation. Cross-attention and relevance-weighted fusion modules were proposed to better align retrieved evidence with generation, improving factual grounding and reducing redundancy [7], [11]. Research also expanded toward multimodal and hierarchical retrieval settings, incorporating structured metadata and non-textual sources alongside text to address increasingly complex information demands [6][16].

Despite these advances, traditional RAG architectures remain limited by their flat retrieval paradigm. Treating document chunks as independent units restricts the modelling of inter-document relationships and multi-hop reasoning, as relational structure is not explicitly encoded in the retrieval process [4][8]. These limitations have directly motivated the development of structured retrieval paradigms such as GraphRAG and knowledge graph enhanced RAG, representing a shift toward relationally informed context construction [3][19].

D. Trend of GraphRAG Research

Graph-augmented retrieval architectures, commonly referred to as GraphRAG, represent a structural extension of traditional RAG in which retrieval is guided by explicit semantic relationships rather than flat similarity alone [3][8]. Instead of retrieving independent chunks, GraphRAG frameworks organise knowledge into graph structures, with nodes representing semantic units such as entities or sentences and edges encoding relationships including co-occurrence, hierarchy, or semantic proximity [3][13]. This structured representation enables multi-hop reasoning and the modelling of long-range dependencies that are typically inaccessible to flat retrieval systems [4][8]. GraphRAG research has concentrated on three core areas. The first is graph construction and indexing, where documents are transformed into graph representations that capture both local content and global relational structure [13][19]. Through entity extraction and relation inference, nodes can encode semantic linkages reflecting real-world knowledge hierarchies, allowing retrieval to consider paths and connections rather than isolated similarity scores [13][19]. The second area is graph-guided retrieval, in which queries traverse graph topology using neighbourhood expansion, community detection, or path scoring to retrieve structurally relevant evidence across multiple hops [4][6][22]. These methods are particularly suited to tasks requiring integrated evidence from multiple documents, such as multi-document QA and long-context summarisation [5][8].

The third area explores graph integration at the generation stage, where retrieved subgraphs inform attention mechanisms or intermediate reasoning steps, reducing hallucinations and improving logical coherence [7][8]. Query-centric graph designs, for example, introduce synthetic query nodes as abstractions between raw text and entity-level representations, improving retrieval efficiency and interpretability while reducing redundancy [22]. Variants such as PathRAG further refine this approach by optimising relational path selection to minimise noise and token overhead [23].

Overall, GraphRAG reflects a shift from similarity-based retrieval toward relational semantics, where retrieval becomes inherently structural rather than purely vector-driven [3][19]. This enables reasoning over entity relationships and information flows that span multiple documents, positioning GraphRAG as a promising architecture for complex, structured QA scenarios where flat chunk retrieval is insufficient [4][5][20].

While both traditional RAG and GraphRAG have been proposed for retrieval-augmented question answering, it remains unclear to what extent observed performance differences are attributable specifically to retrieval structures when other factors are held constant. In particular, the impact of flat versus graph-based retrieval on semantic alignment, contextual relevance, and grounding behaviour across different language models has not been systematically isolated under identical preprocessing, embedding, and evaluation conditions. This paper addresses these gaps through a controlled comparison of traditional RAG and GraphRAG pipelines using shared retrieval parameters, language model backbones, and evaluation metrics. The following sections describe the experimental methodology, present quantitative results across lexical, semantic, and grounding-based metrics, and discuss the implications of retrieval structure on performance, limitations, and future research directions.

III. METHODOLOGY

The methodology was comprised of the five steps detailed below.

A. Preprocessing and chunking

We segmented each document with Recursive Character Text Split (RCTS), with a chunk size of 1000 and overlap of 30 tokens. This method was selected to ensure contextual

continuity while maintaining manageable retrieval units. The 30-token overlap corresponds to approximately 3% of the total chunk size, which was considered sufficient to preserve cross-boundary semantic continuity between adjacent segments. Larger overlaps would increase redundancy and token overhead during retrieval and generation, potentially affecting efficiency. Given the relatively large 1000-token chunk size, a smaller proportional overlap was deemed reasonable compared to higher-percentage overlaps typically used for shorter text segments

B. Vectorization

Each document chunk was transformed into a dense vector representation using *nomic-embed-text*, a transformer-based text embedding model designed to capture semantic similarity in high-dimensional continuous space. The model maps different lengths of text inputs into fixed-size embeddings, that semantically related chunks are positioned closer together under cosine similarity, enabling effective nearest-neighbour retrieval. The embedding step is the foundation for downstream retrieval by encoding both lexical content and higher-level semantic relationships, allowing conceptually aligned statements from the document to be retrieved even when surface wording differs. Dense vectorization is important for the dataset used in this experiment, where paraphrasing and domain-specific terminology are common and exact keyword overlap is insufficient. The resulting embeddings were indexed in a vector database to support efficient similarity search during inference.

C. Preprocessing and chunking Retrieval

For the traditional RAG model, the retriever selected the top-*k* text chunks based on cosine similarity. While for the GraphRAG, the retriever performed neighbourhood exploration through document-level graph, where the nodes represented semantically related sentences, and the edges represented cosine similarity relationships.

D. Language Models and Answer Generation

To isolate the effect of retrieval structure while controlling for generative capacity, two large language models hosted via Ollama were used consistently across both RAG and GraphRAG pipelines.

The first model, LLaMA 3.2 3B, is a lightweight decoder-only transformer with approximately 3 billion parameters. It uses Grouped-Query Attention for efficient inference and has been trained on a large multilingual corpus with instruction tuning to improve reasoning and dialogue quality. While compact, it provides strong baseline performance in retrieval-augmented settings and is well suited for evaluating scenarios where retrieval quality is the primary performance bottleneck. However, its limited parameter count, and shorter effective reasoning depth may constrain performance on tasks requiring complex compositional inference [15].

The second model, Phi-4, is a reasoning-optimised decoder-only transformer developed by Microsoft, trained on high-quality curated data and synthetic textbook-style reasoning corpora. With a larger parameter count and a 16k-token context capacity, Phi-4 is better suited for structured inference and multi-hop reasoning. This makes it particularly effective in GraphRAG pipelines, where graph-structured retrieval provides richer contextual signals that can be more fully exploited by a higher-capacity model. Although computationally heavier, Phi-4 has been shown to exhibit stronger faithfulness and contextual alignment in question answering tasks [16].

For both models, retrieved context was concatenated and passed to the language model using a structured prompt template, ensuring consistent generation conditions across experimental settings. Both pipelines used nomic-embed-text for embedding and similarity computation to avoid embedding-induced bias.

E. Preprocessing and chunking Retrieval Evaluation

The generated answers were compared with the ground truth reference answers using both lexical (BLEU, ROUGE) and semantic (RAGAS) metrics. To comprehensively assess RAG performance on open-ended question answering, we adopted a multi-dimensional evaluation framework combining lexical overlap metrics, semantic similarity measures based on vector space modelling [29], and retrieval-grounding metrics [28]. This choice reflects the need to evaluate not only surface-form similarity, but also meaning preservation, factual grounding, and query alignment, which are critical in government policy QA where correct paraphrasing is often preferable to verbatim reproduction.

1) BLEU (Bilingual Evaluation Understudy): BLEU is a precision-oriented n-gram overlap metric used for automatic evaluation of machine translation [26]. It measures lexical similarity between a generated answer and a reference text. It is defined as:

$$\text{BLEU} = \text{BP}.\exp\left(\sum_{n=1}^{N} w_n \, log p_n\right) \quad (1)$$

where p_n is the modified n-gram precision, w_n are uniform weights, and BP is the brevity penalty. BLEU is included to ensure comparability with prior RAG literature but is expected to yield low scores due to the abstractive and paraphrastic nature of responses. The brevity penalty BP is defined as:

$$BP = \begin{cases} 1, & c > r \\ \exp\left(1 - \frac{r}{c}\right), & c \leq r \end{cases} \quad (2)$$

where:

- c is the total length (in tokens) of the generated answer, and
- r is the total length (in tokens) of the reference answer.

2) ROUGE (Recall-Oriented Understudy for Gisting Evaluation): ROUGE is a recall-oriented evaluation metric commonly used for summarisation and long-form text generation [27]. Unlike BLEU, which emphasises precision, ROUGE measures how much of the reference content is

covered by the generated answer. In this study, ROUGE-1, ROUGE-2, and ROUGE-L were used to capture different aspects of content overlap.

ROUGE-*N* measures *n*-gram recall and is defined as:

$$ROUGE - N = \frac{\sum_{g \in Ref} \min(Count_{gen}(g), Count_{ref}(g))}{\sum_{g \in Ref} Count_{ref}(g)} \quad (3)$$

where:

- g denotes an *n*-gram,
- G_N(Ref) is the set of all *n*-grams of length *N* appearing in the reference answer,
- $Count_{ref}$(g) is the number of times *n*-gram g appears in the reference text, and
- $Count_{gen}$(g) is the number of times *n*-gram g appears in the generated answer.

ROUGE-1 corresponds to unigram (N=1) overlap and primarily captures coverage of individual content words and key terms. ROUGE-2 corresponds to bigram (N=2) overlap and captures short phrase-level consistency and local fluency.

ROUGE-L differs fundamentally from ROUGE-1 and ROUGE-2 in that it does not rely on fixed-length *n*-grams. Instead, it measures the Longest Common Subsequence (LCS) between the generated answer and the reference text. The LCS captures the longest sequence of tokens that appear in both texts in the same order, though not necessarily contiguously. As a result, ROUGE-L is more sensitive to global sentence structure and discourse-level coherence, making it particularly suitable for evaluating longer, abstractive answers where key ideas may be reordered or paraphrased.

3) Cosine Similarity (Semantic Similarity): Cosine similarity is a standard metric in vector space information retrieval models [29]. It evaluates semantic alignment between generated answers and reference texts using embedding representations:

$$\cos(\theta) = \frac{u.v}{||u||\,||v||} \quad (4)$$

where u and v are the embedding vectors of the generated and reference texts. This metric captures meaning equivalence independent of surface form and serves as the primary indicator of semantic fidelity in abstractive dataset QA**.**

4) Faithfulness (RAGAS): Faithfulness measures whether claims in the generated answer are supported by retrieved context rather than hallucinated. Using the RAGAS framework, faithfulness is computed as the proportion of answer statements that can be grounded in retrieved evidence. This metric is critical for policy QA, where evidence-backed responses are required.

5) Context Relevance (RAGAS): Context relevance measures how directly the generated answer addresses the user query, independent of factual correctness. It quantifies semantic alignment between the query and the answer and penalises responses that are generic or tangential, providing insight into retrieval quality and query focus.

6) Metric Selection Rationale: Together, these metrics form a complementary evaluation suite:

- BLEU and ROUGE assess lexical overlap and content coverage.
- Cosine similarity captures semantic equivalence.
- Faithfulness evaluates grounding and hallucination control.
- Context relevance measures query responsiveness.

This multi-metric approach is necessary because no single metric adequately captures the performance of retrieval-augmented systems on open-ended, policy-driven QA tasks. In particular, reliance on lexical metrics alone would obscure meaningful improvements in semantic grounding and retrieval structure precisely, the aspects that GraphRAG is designed to enhance.

IV. PERFORMANCE EVALUATION

The evaluation was conducted on a policy-oriented question answering dataset derived from the dh-gen-ai-intensive-course-capstone-2025q1 corpus. The dataset consists of eight government coalition policy documents covering national reforms across governance, digital policy, economic growth, education, foreign policy, migration, and democratic stability. These documents are long-form, concept-dense, and highly abstractive, making them a challenging benchmark for retrieval-augmented generation. A total of 19 open-ended analytical questions were designed to probe factual recall, semantic understanding, multi-document integration, and policy reasoning. Two retrieval paradigms were evaluated: traditional RAG and GraphRAG. Each paradigm was tested under two language models namely LLaMA 3.2 (3B) and Phi-4.

All experiments used identical embeddings (nomic-embed-text), chunking strategy (RCTS), chunk size (1000 tokens), and overlap (30 tokens), ensuring that observed differences are attributable to retrieval structure and model behaviour rather than preprocessing variance.

To contextualise the results, it is important to note that lexical metrics such as BLEU and ROUGE typically yield modest values in open-ended, abstractive QA tasks, particularly when multiple valid phrasings exist. BLEU scores below 0.15 are common in long-form generation, while ROUGE-1 scores in the range of 0.20 - 0.30 generally indicate meaningful content overlap. In contrast, cosine similarity values above 0.65 suggest strong semantic alignment in embedding space. RAGAS-based metrics (faithfulness and context relevance) are relative measures rather than absolute quality indicators and are most informative when comparing systems under identical experimental conditions.

The results, using the various metrics, are presented in Table 1. BLEU scores are consistently low across all configurations, ranging from approximately 0.002 to 0.12, which is expected for open-ended, abstractive policy QA. BLEU is highly sensitive to exact wording and penalises paraphrasing, synonym usage, and sentence restructuring. Across both LLaMA 3B and Phi-4, traditional RAG

marginally outperforms GraphRAG on average BLEU, suggesting that flat retrieval occasionally reproduces phrasing closer to the reference text. However, this does not imply superior answer quality, as policy responses typically require abstraction and synthesis rather than verbatim reproduction. These observations are consistent with prior findings showing that BLEU correlates poorly with semantic correctness and factual adequacy in abstractive question answering and retrieval-augmented generation systems, where multiple valid phrasings exist for a given answer [12][13].

TABLE I. AVERAGE PERFORMANCE

Metric	RAG (Llama 3B)	GraphRAG (Llama 3B)	RAG (Phi4)	GraphRAG (Phi4)
BLEU	0.0278	0.0247	0.0358	0.0273
ROUGE-1	0.2128	0.2283	0.2540	0.2283
ROUGE-2	0.0715	0.0729	0.0816	0.0748
ROUGE-L	0.1586	0.1569	0.1808	0.1629
Cosine Similarity	0.6776	0.6925	0.6368	0.6847
Faithfulness (RAGAS)	0.3085	0.2866	0.2887	0.3534
Context Relevance	0.2957	0.3093	0.2426	0.3263

ROUGE metrics provide a more informative signal than BLEU for this task. Across all models and we can see that GraphRAG consistently matches or outperforms RAG on ROUGE-1 and ROUGE-2 and ROUGE-L improvements are modest but systematic. For example, in questions requiring multi-section integration (e.g., digital sovereignty, international cooperation, cybersecurity), GraphRAG demonstrates higher ROUGE-1 and ROUGE-L scores, indicating better recall of salient policy concepts and more coherent structural alignment. This suggests that GraphRAG retrieves a more complete and topically coherent subset of evidence, even when exact phrasing differs from the ground truth.

Cosine similarity emerges as the most discriminative metric for this evaluation. Across nearly all questions, we note that GraphRAG achieves higher semantic similarity than RAG and the improvement is especially pronounced with Phi-4. Using LLaMA 3B, GraphRAG improves average cosine similarity from 0.6776 to 0.6925, while using Phi-4 the improvement is even larger (0.6368 to 0.6847). These values fall well within the range expected for high-quality abstractive QA, indicating that GraphRAG better preserves the meaning of policy answers even when surface wording diverges. This result directly reflects the structural advantage of GraphRAG, which retrieves semantically linked evidence paths rather than isolated text chunks.

For the Faithfulness measure, results show a model-dependent interaction: with LLaMA 3B, traditional RAG slightly outperforms GraphRAG; with Phi-4, GraphRAG shows a substantial improvement, achieving the highest overall faithfulness score (0.3534). This indicates that GraphRAG benefits more strongly from stronger reasoning-oriented models, which are better able to exploit structured retrieval signals and relational evidence during generation.

Context relevance consistently favours GraphRAG across both models, with the largest gains observed when using Phi-4. GraphRAG achieves an average relevance score of 0.3263, compared to 0.2426 for RAG. This confirms that GraphRAG's structural retrieval mechanism helps suppress retrieval noise and maintain tighter alignment between the user query and the generated response an essential property for policy QA where tangential accuracy is insufficient.

Overall, several key patterns emerge from the results:

- Lexical metrics underestimate GraphRAG performance, particularly for abstractive answers.
- Semantic and grounding metrics consistently favour GraphRAG, especially when paired with stronger models.
- Phi-4 amplifies GraphRAG's advantages, suggesting that structured retrieval and advanced reasoning capabilities are complementary.
- Flat RAG remains competitive on surface overlap, but struggles with semantic integration, relevance control, and factual grounding.

These findings support the hypothesis that GraphRAG represents a structural evolution of RAG, particularly well-suited for complex, multi-document, policy-driven question answering.

A. Limitations

The primary limitations observed in this study arise from the interaction between model capacity, retrieval structure, domain specificity, and evaluation methodology. Both traditional RAG and GraphRAG operate over policy documents that are inherently abstractive, normative, and semantically dense, which substantially reduces lexical overlap with reference answers and leads to systematically low BLEU and moderate ROUGE scores despite semantically correct outputs. Traditional RAG is further constrained by its flat retrieval paradigm, which treats document chunks as independent units and therefore fails to model inter-document relationships or multi-hop reasoning paths that are common in coalition policy questions. GraphRAG mitigates this issue through structured retrieval but introduces its own limitations, particularly sensitivity to graph construction quality, entity linking noise, and traversal depth selection, which can propagate weakly relevant nodes into the retrieved context and slightly depress faithfulness scores. Across both approaches, the use of general-purpose embedding models and relatively small language models limits fine-grained representation of policy-specific terminology and long-range reasoning, while reliance on single-reference lexical metrics underestimates answer quality by penalising valid paraphrasing, synthesis, and

abstraction that are essential for policy-oriented question answering.

V. CONCLUSION AND FUTURE WORK

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper.

While both retrieval-augmented approaches exhibit limitations under traditional lexical evaluation metrics, the results demonstrate that GraphRAG provides a structurally superior and more future-proof framework for complex, policy-oriented question answering. Traditional RAG remains competitive in constrained scenarios where surface-level retrieval suffices, but its flat architecture limits multi-hop reasoning and relational synthesis across documents. GraphRAG consistently achieves stronger semantic alignment and query relevance, particularly when paired with more capable language models, and its remaining weaknesses are largely attributable to implementation maturity rather than conceptual design. Overall, the evidence supports GraphRAG as the more effective and scalable architecture for high-stakes, knowledge-intensive applications where correctness, grounding, and interpretability outweigh surface-form similarity.

Future work will focus on improving graph construction precision, node weighting, and dynamic subgraph selection to reduce retrieval noise, as well as exploring hybrid retrieval strategies and adaptive chunking to improve evidence coverage. Additional gains may be achieved through embedding fine-tuning, planning-aware and citation-constrained generation, and the use of models with stronger long-context reasoning capabilities. Finally, future evaluation should prioritise semantic and grounding-based metrics with multiple reference answers to better reflect the abstractive nature of policy QA, as low BLEU and moderate ROUGE scores primarily reflect metric mismatch rather than model failure.

ACKNOWLEDGMENT

This research is funded by UKRI (UK Research and Innovation) under the Hartree National Centre for Digital Innovation Programme (Hartree Hub Northern Ireland - Home.)

REFERENCES

[1] P. Lewis, et al., “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 9459–9474, 2020.

[2] G. Izacard and E. Grave, “Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering,” arXiv preprint arXiv:2007.01282, 2020.

[3] M. Yasunaga, X. Ren, B. Bosselut, P. Liang, and J. Leskovec, “GraphRAG: Retrieval-Augmented Generation with Graphs,” arXiv preprint arXiv:2501.00309, 2025.

[4] Y. Zhao, Z. Wang, H. Chen, and J. Li, “When to Use Graphs in Retrieval-Augmented Generation: A Study on Complex Question Answering,” arXiv preprint arXiv:2506.05690, 2025.

[5] D. Zerva, K. Kapanipathi, V. Karpukhin, and S. Riedel, “GraphRAG-Bench: A Benchmark for Graph-Based Retrieval-Augmented Generation,” arXiv preprint arXiv:2506.02404, 2025.

[6] J. Lin, Z. Wang, Y. Li, and X. Ren, “From Local to Global: A Graph-RAG Approach to Query-Focused Summarization,” arXiv preprint arXiv:2404.16130, 2024.

[7] K. Shuster, S. Poff, M. Chen, D. Kiela, and J. Weston, “Retrieval Augmentation Reduces Hallucination in Conversation,” in Proc. 2021 Conf. on Empirical Methods in Natural Language Processing (EMNLP), pp. 3784–3803, 2021.

[8] H. Zhang, Y. Li, J. Gao, and Z. Liu, “RAG vs. GraphRAG: A Systematic Evaluation and Key Insights,” arXiv preprint arXiv:2502.11371, 2025.

[9] Deep-PolyU Research Group, “Awesome GraphRAG,” GitHub repository. [Online]. Available: https://github.com/Deep-PolyU/Awesome-GraphRAG, 2025.

[10] V. Karpukhin, et al., “Dense Passage Retrieval for Open-Domain Question Answering,” in Proc. 2020 Conf. on Empirical Methods in Natural Language Processing (EMNLP), pp. 6769–6781, 2020.

[11] L. Gao, X. Ma, J. Lin, and Z. Liu, “Rethink Training of Retrieval-Augmented Generation for Open-Domain Question Answering,” in Proc. 60th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 2061–2077, 2022.

[12] R. Zhong, M. Yasunaga, and X. Ren, “Extractive Retrieval-Augmented Generation for Policy and Regulatory Documents,” arXiv preprint, 2023.

[13] M. Yasunaga, H. Chen, Y. Ren, and P. Liang, “Graph-Based Retrieval and Reasoning for Open-Domain Question Answering,” in Proc. 59th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 163–175, 2021.

[14] Meta AI, “LLaMA 3.2-3B,” Hugging Face Model Repository. [Online]. Available: https://huggingface.co/meta-llama, 2024.

[15] Microsoft, “Introducing Phi-4: Microsoft’s Newest Small Language Model Specializing in Complex Reasoning,” Microsoft Community Hub, 2024.

[16] S. Gupta, R. Ranjan, and S. N. Singh, “A Comprehensive Survey of Retrieval-Augmented Generation: Evolution, Current Landscape, and Future Directions,” arXiv preprint, 2024.

[17] A. Gan, et al., “Retrieval-Augmented Generation Evaluation in the Era of Large Language Models: A Comprehensive Survey,” arXiv preprint, 2025.

[18] Y. Huang and J. Huang, “A Survey on Retrieval-Augmented Text Generation for Large Language Models,” arXiv preprint, 2024.

[19] B. Peng, Y. Zhu, Y. Liu, X. Bo, and H. Shi, “Graph Retrieval-Augmented Generation: A Survey,” arXiv preprint, 2024.

[20] Q. Zhang, et al., “A Survey of Graph Retrieval-Augmented Generation for Customized Large Language Models,” arXiv preprint, 2025.

[21] QualityPoint Technologies, “The Evolution of Retrieval-Augmented Generation,” QualityPoint Technologies Blog, 2025.

[22] Emergent Mind Research, “Graph-Centric RAG Frameworks and Query-Centric Design Studies,” Emergent Mind Overview, 2025.

[23] Research Community, “PathRAG: Path-Focused Graph Retrieval-Augmented Generation Innovations,” Community Research Summary, 2025.

[24] C. Callison-Burch, M. Osborne, and P. Koehn, "Re-evaluating the role of BLEU in machine translation research," in Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Trento, Italy, 2006, pp. 249–256.

[25] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating text generation with BERT," in Proceedings of the 8th International Conference on Learning Representations (ICLR), Addis Ababa, Ethiopia, 2020.

[26] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Philadelphia, PA, USA, 2002, pp. 311–318.

[27] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. ACL Workshop Text Summarization Branches Out*, Barcelona, Spain, 2004, pp. 74 - 81.

[28] S. Es, S. James, A. Espinosa-Anke, and S. Schockaert, "RAGAS: Automated evaluation of retrieval augmented generation," arXiv:2309.15217, 2023.[retrieved: Feb. 2026].

[29] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. New York, NY, USA: McGraw-Hill, 1983.

[30] Mindful Matrix, "Building LLM application using RAG" Substack, [Online]. Available: https://mindfulmatrix.substack.com/p/build-a-simple-llm-application-with, 2024.

[31] D. vonThenen, "Beyond Vectors – Knowledge Graphs & RAG Using Gradient," DigitalOcean Community, [Online]. Available: https://www.digitalocean.com/community/tutorials/beyond-vectors-knowledge-graphs-and-rag, 2025.

Proposal of a Semi-Automatic Classification Method for Estimating Conceptual Understanding in Short Answer Grading for Semi-Open-Ended Questions Using Word Co-occurrence Networks

Katsuko T. Nakahira
Nagaoka University of Technology
Nagaoka, Niigata, Japan
Email: katsuko@vos.nagaokaut.ac.jp

Muneo Kitajima
Nagaoka University of Technology
Nagaoka, Niigata, Japan
Email: mkitajima@kjs.nagaokaut.ac.jp

***Abstract*— One of the effective method for estimating learners' level of understanding of acquired concepts involves using semi-open-ended questions. The method is well known to be particularly beneficial for questions requiring scientific explanations, and it is adopted in large-scale academic achievement tests and trend assessments. On the other hand, Short Answer Grading often relies on manual scoring by markers, raising concerns about workload and the depth of insight into specialized knowledge. To alleviate marker workload, automated evaluation combining natural language processing and machine learning, or utilizing generative AI, is anticipated. However, introducing machine learning requires large amounts of training data and also faces issues related to the native language of the test takers. As one solution to these problems, a method that enables classifying learners' understanding levels with minimal effort based on collected short answers, even under conditions of limited information for machine learning, is also considered beneficial. In this paper, we propose a method that classifies short answers into three levels based on conceptual understanding depth as one approach to short answers grading for semi-open-ended questions. The proposed method applies degree analysis, well-known in word co-occurrence networks.**

Keywords-Semi-Open-Ended question; short answer grading; word co-occurrence network; degree analysis.

I. Introduction

With the advancement of natural language processing and generative Artificial Intelligence (generative AI), many researches in recent years have focused on the automated evaluation of descriptive responses to open-ended question.

For questions requiring scientific explanations, it is preferable to use a Short Answer format for semi-open-ended problems, and the marker should grade them manually. In this paper, the term "semi-open-ended" answer refers to response that, while open-ended, are subject to certain constraints on the perspective taken—such as an interpretation of trends in a statistical graph. While the usefulness of large-scale academic achievement tests and trend tests has been noted, the burden on markers due to the enormous number of examinees and the depth of insight required for specialized knowledge have long been pointed out. This marker-based grading, known as Short Answer Grading (SAG), involves assessment to measure student learning outcomes. However, in today's world where student numbers or their growth increases exponentially, this task has become a difficult challenge for educators. To address the above issues, Automated SAG (ASAG) systems combining natural language processing and machine learning, as well as ASAG systems utilizing generative AI, have been implemented.

Analysis of text sentiment in natural language processing has long been integrated with knowledge from network science, yielding various practical studies in the form of Word Co-occurrence Networks (WCNs). In recent years, research into the structure and dynamics of complex networks has attracted the attention of both academic researchers and practitioners.

Applications of the WCN method to long text include the following examples. Some researcher created WCNs for academic papers published in the field of information bibliometrics (see Sedighi [1]) or bioinformatics (see Li et.al. [2]) and analyzed their structural features and depth. These studies confirmed small-world properties and provided insights into temporal transitions of concepts and word usage. Additionally, Liang et al. [3] constructed WCNs for standard English articles, calculated the spectra of their adjacency matrices, and identified characteristic distributions in the spectral density distribution through qualitative analysis of the spectra.

Research applying WCN to short text corpora includes, for example: Garg et al.'s [4] application to microblogs, Fudolig et al.'s [5] application to political tweets, and Amancio et al.'s [6] application to short texts. In each case, the structure of WCNs differed from that of longer texts, suggesting the existence of characteristic parameters and the necessity of specific analytical methods.

Additionally, there is a research example focusing on the context of text authors, Millington and Luz [7]. Using the ADReSS Challenge dataset, a natural speech dataset, they constructed WCNs from conversation records of healthy individuals and potential Alzheimer's disease patients and compared structural metrics. The results revealed distinct features in heterogeneity, centrality, and edge density specific to the networks of Alzheimer's patients, with many network characteristics being driven by word frequency. This suggests that, overall, classification between the healthy control group and the Alzheimer's group is possible based solely on the structure of the WCN.

Proposals for ASAG systems combining natural language processing and machine learning include, for example, the following: A method for separating essay groups into subsets representing similar graders using explanatory variables and clustering (Zupanc and Bosnić [8]), a method utilizing rubric

 ISBN: 978-1-68558-375-0

information (Wang et al. [9]), a method that introduces a Transformer encoder layer into a BERT model and trains its weights solely on the text of the rubric criteria table (Condor et al. [10]), focusing on semi-open-ended short-answer questions and integrating both general domain knowledge and domain-specific information (Zhang et al. [11]), and one that utilizes very basic natural language processing techniques such as N-grams and word embedding technology along with machine learning algorithms (Lertchaturaporn and Pokpong [12]), among others. Additionally, there are investigations into the effects on learners' trust in grading and learning motivation (Conijin et al. [13]). At present, ASAG systems centered on natural language processing and machine learning are mainstream, though there are many applications of ASAG systems using generative AI (Jamil and Hameed [14], Chang and Ginter [15], Grévisse [16], Mello et al. [17]).

Taking these trends in consideration, future possibilities for ASGA include not only focusing on accurate grading but also categorizing response levels and conducting detailed reviews as needed for assigned levels. Particularly when using machine learning, the one of greatest effort lies in the necessity of training the system using past answer patterns and their corresponding grading results. Additionally, working language problem adopting toward machine learning issues arise, especially for educational institutions that accept many international students. Of course, to reduce the effort required for grading, the use of machine learning or generative AI is ultimately desirable. However, as an alternative approach, even if it doesn't reduce effort as much as machine learning, a mechanism that allows classifying learners' level of understanding with minimal effort based on collected short answers—even in situations where information for machine learning is scarce—is worth considering. Based on this idea, we propose a framework for educational settings to classify learners' short answers to semi-open-ended questions—even with minimal knowledge—into three levels: 1) Simple recall, 2) Explanation through reconstruction of learned knowledge, 3) Application of learned knowledge. Through this framework, there is a possibility to categorize learners' comprehension levels and knowledge expression during lectures or briefings, thereby supporting subsequent educational efforts.

This paper is structured as follows. Section 2 describes the estimation of WCN structures based on the SAG process for semi-open questions, using the concept of degree in complex networks. Section 3 introduces the deep structure of the nearest-neighbor plane based on the framework described in Section 2. It then discusses key parameters for estimating this deep structure by performing a simplified simulation on it. We conclude the paper in Section IV.

II. Short Answer Generating Process for Semi-Open-Ended Question and Word Co-Occurrence Network

A. Selecting the Words for Generating Answer for Semi-Open-Ended Question

Examples of using semi-open-ended questions include taking minutes after a lecture or reading comprehension exercises when figures or tables are presented. A common purpose for adopting semi-open-ended questions is to assess learners' level of understanding when they possess specific prior knowledge. Here, we describe the cognitive model involved in solving semi-open-ended questions, specifically the process of generating semi-open-ended answers.

First, learners acquire newly presented concepts as knowledge. This knowledge is, in principle, encoded in some form. In the context of school learning, concepts are fundamentally encoded as language. These encoded concepts are then stored in long-term memory as either technical terms used in specific contexts, general terms used to explain those technical terms, or linked to other technical terms. In this process, the language used to explain the concept is stored in long-term memory by linking several words together.

This shares similarities with the micro structure of Generic Skills introduced by Nakahira et al. [18]. That paper introduced a framework where Generic Skills consist of three stages: perception of physical contents provided through lectures or experiments, collection of perceived contents, and reconstruction. Collected contents become information, and when information is organized, it becomes knowledge. The series of activities suggested that learner behavior differs depending on whether the acquisition action is passive or active. Specifically, it was introduced that in passive acquisition, transitioning collected content beyond simple memorization is difficult, whereas in active acquisition, further exploration of the obtained information and application to similar situations enable transition to the re-representation of acquired knowledge.

In this paper, we apply a series of micro structures to the act of solving semi-open-ended questions. Short answers observed when solving semi-open-ended questions regards as reflecting one of the following three states, depending on the level of mastery of the learned concepts.

- **Level A: Simple Memorization of Learned Knowledge**
  Reciting learned concepts verbatim (simple memorization)
- **Level B: Explanation through reconstruction of learned knowledge**
  Able to explain learned concepts by substituting similar words for the original terms (paraphrasing, somewhat deeper understanding)
- **Level C: Application of learned knowledge**
  Able to use learned concepts to explain similar concepts or phenomena (application of knowledge)

Hence, the vocabulary used in the short answers written in the responses can be also considered to reflect the above proficiency levels.

The phrases selected for short answer writing can be observed as multiple words appearing simultaneously, i.e., in the form of word co-occurrence. In the field of natural language processing, this has traditionally been addressed using Word Co-occurrence Networks (WCNs). This paper adapts this idea, proposing that the selection of word groups used in short answer descriptions and learners' proficiency levels regarding

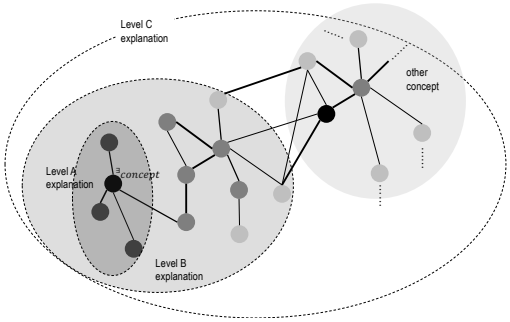


Figure 1. Conceptual diagram of the network of words used to explain a certain concept.

learning concepts are reflected in WCN features. We also propose analyzing this using the following framework.

B. Description of the Relationship between Degree and Conceptual Proficiency Levels for WCN Words Appearing in Short-Answer Questions

For a given word W_i that we adopt when generating text, we set the following assumption. The contexts in which W_i is used can be divided into two categories:

- General words, which are used without reference to a specific context.
- Specific words, which are only used in specific contexts.

Now, we imagine that someone are creating a short answer for a given semi-open-ended question. Since the question is a semi-open-ended given in a proficiency assessment setting, the short answer will contain a mix of two types of descriptions: the precise knowledge expected and descriptions that further develop that knowledge through reasoning. In this process, we assume that descriptions of the concept will be formed through the following steps, depending on the level of understanding (proficiency level) of the knowledge acquired in the past.

- **Level A**
  The concepts learned are explained using specialized words, so terms not commonly used in everyday life are frequently employed. In other words, specific words appears frequently.
- **Level B**
  After simple memorization, the learned knowledge is explained by reconstructing the concepts. Therefore, even when using technical terms, they are simultaneously replaced with alternative expressions. The words adopted at this time are not necessarily technical terms. That is, general words appear mixed among specific words.
- **Level C**
  When explaining similar concepts or phenomena, the application is not necessarily specialized, so technical terms are used alongside general words in short answers. That is, specific words appear among many general words.

The selection of words used to explain a concept depends on the words associated with that concept. This conceptual diagram is shown in Figure 1. When first learning a concept, one can only explain it using the exact term taught. Therefore, explanations will likely be limited to words found within the very narrow scope of Level A. As understanding of the concept broadens slightly, it becomes possible to replace the explanation with words from Level B, using words that can connect to areas outside Level A. With further deepening of understanding, it becomes possible to use words connecting to other concepts to provide explanations that link the concepts together. This state is considered to be Level C.

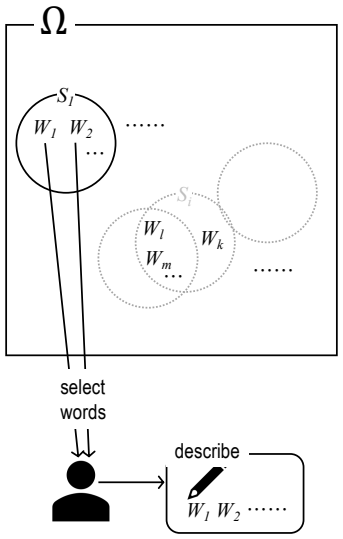


Figure 2. The relation between words selection and short answer writing.

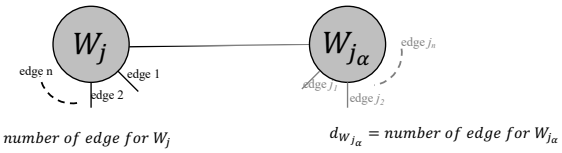


Figure 3. Relationship between nodes and edges for W_j and W_{j_α}.

When considering the "word network" during generating short answer, the process follows the flow described above, and this process can related to human cognitive behavior. Figure 2 shows a conceptual diagram of the word selection process when learners write short answers after being presented with a semi-open-ended question. In the case of semi-open-ended questions administered after a lecture, they are typically used by instructors to assess learners' proficiency levels. Therefore, it is assumed that semi-open-ended questions are set to multiple learners, typically tens to hundreds of individuals. After the lecture concludes, or in some other context, when an instructor provides a semi-open-ended question to learners, learner i creates and submits a short answer S_i. As a result, after the semi-open-ended question is completed, short answers $\boldsymbol{S}$ from n_p respondents can be collected.

$$\boldsymbol{S} = \{S_i(\boldsymbol{W}_i(W_j)) \mid i = 1, \cdots, n_p,\ j = 1, \cdots, n_{w_i}\}.$$

S_i contains n_{w_i} words, and learners express concepts learned by linking multiple W_j.

The number of words W_j can link, i.e., the number of edges when W_j is a node, is called the degree which is denoted by d_{W_j}. The relationship between W_j, d_{W_j}, and the associated

TABLE I. TREND OF DEGREE PAIRS $(d_p,\ d_q)$ FOR WORD PAIRS $W_p,\ W_q$. "C" DENOTES COMPLEX, "A" DENOTES AVERAGE, "S" DENOTES SPARSE.

		degree for word p		
		large	moderate	small
degree for word q	large	c/c	a/c	s/c
	moderate	c/a	a/a	s/a
	small	c/s	a/s	s/s

W_{j_α} and $d_{W_{j_\alpha}}$ is shown in Figure 3. In this context, d_{W_j} can be interpreted as follows: if W_j is a global word, it can link to many words, meaning large d_{W_j}; if it is a specific word, it can link only to a few words, meaning small d_{W_j}. Therefore, this situation suggests that by performing degree analysis on W_j, it may be possible to investigate a learner's proficiency level regarding specific concepts they have learned, even before confirming the content of the text. This is described as follows.

Let $\mathbf{\Omega}$ denote the set of all unique words W_α appearing in $\boldsymbol{S}$.

$$\mathbf{\Omega} = \{W_\alpha \mid \alpha = 1, \cdots, n_{w_t}\}$$

Here, n_{w_t} is the number of the set. Each W_α take a value of d_{W_α}.

When the learners generate short answers for semi-open-ended questions, they extract n_{w_i} words from $\mathbf{\Omega}$ and concatenate them in an appropriate order. In this case, the word(s) linked to W_j each possess a degree. To grasp the general trend before reading the set of short answers, examining the degree of the word(s) linked to W_j allows for a simple check, as shown in Table I. Table I shows that when considering a word pair $(W_p,\ W_q)$, the combination of the degrees of the degree pair $(d_p,\ d_q)$ results in expressions ranging from c/c (simple explanations using words not tied to specific contexts) to s/s (complex explanations using words tied to specific contexts).

Since W_j can connect to d_{W_j} words, there are d_{W_j} possible combinations of $(W_p,\ W_q)$. Given this characteristic, it would be beneficial to have a method that simultaneously represents the properties of W_j and the properties of all words linked to W_j. As an idea, treating the number of words W_j itself can link to and the total number of words connected to W_j that can link to other words as features would enable efficient analysis. We describe the method as follows.

d_{W_j} represents the number of words W_j can link to other words, or in other words, the number of first-order nearest neighbors. This is denoted as $d_{j_{NN_1}}$. $d_{j_{NN_1}}$ suggests it can serve as a criterion for determining whether W_j is a general word or a specific word.

Next, consider the characteristics of the words $W_{j_\alpha}(\alpha = 1, 2, \cdots, d_{W_j})$ connected to W_j. Like W_j, W_{j_α} also has degree d_{j_α}. The property of words that can link to W_j, namely their degree of general/specific words, is expressed as the total number of nodes linked to W_j that possess d_{j_α}. This corresponds to determining the number of second-order nearest neighbors from the perspective of W_j. We denote this as $d_{j_{NN_2}}$, and described by the following equation.

$$d_{j_{NN_2}} = \sum_{\alpha=1}^{d_{j_{NN_1}}} (d_\alpha - 1) \tag{1}$$

The reason for subtracting 1 from d_α is to prevent double counting of connections, since W_{j_α} is originally linked to W_j.

The $(d_{NN_1},\ d_{NN_2})$ calculated as described above are represented numerically, allowing their occurrence frequency to be examined. If the occurrence frequency of a specific $(d_{NN_1},\ d_{NN_2})$ is high, we can consider that $\boldsymbol{S}$ is expressed as a phrase possessing a specific combination of word properties. Therefore, it suffices to find a method to reproduce the distribution of $(d_{NN_1},\ d_{NN_2})$.

III. REPRODUCTION OF DEGREE DISTRIBUTION THROUGH SUPERPOSITION OF PROBABILITY DENSITY DISTRIBUTIONS

The frequency distribution of $(d_{NN_1},\ d_{NN_2})$ shown in the previous chapter is the distribution of WCN pairs NN_1 and NN_2 derived from human-generated text. There are many various classification methods and metrics within complex networks to investigate this tendency. This paper posits that the frequency distribution of $(d_{NN_1},\ d_{NN_2})$ follows some probability distribution, as it is generated in accordance with human thought. While numerous types of probability distributions exist, this paper assumes a two-dimensional normal distribution for simplicity.

The two-dimensional normal distribution for variables x and y is described by the following equation.

$$\begin{aligned} p(x,y) &= \frac{1}{R}\exp\left(-\frac{1}{2(1-\rho^2)}(X^2+Y^2-C)\right), \\ &\text{here,} \\ R &= 2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}, \\ X &= \frac{x-\mu_x}{\sigma_x}, \\ Y &= \frac{y-\mu_y}{\sigma_y}, \\ C &= \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}. \end{aligned}$$

When mixing $p(x,\ y)$, let K denote the number of mixture components. The actual distribution $F(d_{NN_1},\ d_{NN_2})$ of $(d_{NN_1},\ d_{NN_2})$ can be described as follows.

$$F(d_{NN_1},\ d_{NN_2}) = \sum_k^K \pi_k p_k(x,\ y),\ \text{here,}\ \sum_k \pi_k = 1$$

For $F(d_{NN_1},\ d_{NN_2})$, considering a specific mixture component k, the following exist for $p_k(x,\ y)$: mean $\mu_x,\ \mu_y$, standard deviation $\sigma_x,\ \sigma_y$, correlation coefficient ρ. These are determined for each NN_1 and NN_2 belonging to k. Furthermore, for each p_k, the mixing weight π_k should ideally also be considered.

Using the above parameters and equations, in this paper we set $K = 2$ to create a template matching the distribution of $F(d_{NN_1},\ d_{NN_2})$. Then we discuss the tendency of $\boldsymbol{S}$ when a distribution resembling this template is observed. Table II shows the parameters we used, and calculations were performed using various combinations. Since p_k includes

TABLE II. THE PARAMETER SET USED FOR COMPUTING THE PROBABILITY DENSITY DISTRIBUTION. THE COMPUTATION RANGE IS FROM -5 TO 5.

parameter	value set
μ_{NN_1}	$-3,\ -1,\ 0.5,\ 1,\ 3$
μ_{NN_2}	$-3,\ -1,\ 0.5,\ 1,\ 3$
σ_{NN_1}	$1,\ 2,\ 3$
σ_{NN_2}	$1,\ 2,\ 3$
ρ	$0.1,\ 0.25,\ 0.5,\ 0.75,\ 0.9$
π_k	0.5

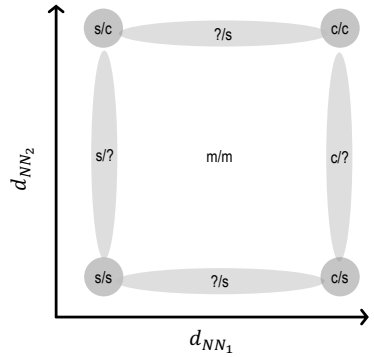


Figure 4. Prediction of the occurrence of each state in Table 1 using mixed synthesis. The gray shaded area indicates the approximate distribution.

$\mu_{NN_1},\ \mu_{NN_2},\ \sigma_{NN_1},\ \sigma_{NN_2},\ \rho$, the parameters for each k were calculated by combining the values from Table II and summing them with each weight π_k set to 0.5. Figure 5, Figure 6, and Figure 7 show the computational results. Based on the results, we discuss about the behavior. All figures display the probability density at its maximum value. When calculating the probability density distribution, the computational range for both NN_1 and NN_2 is from -5 to 5.

A. Distribution of $(d_{NN_1},\ d_{NN_2})$ under the Mixture Model

The template created as described above, which mimics the $(d_{NN_1},\ d_{NN_2})$ distribution, is represented as a density distribution on a plane with d_{NN_1} on the horizontal axis and d_{NN_2} on the vertical axis, as shown in Figure 4. In this paper, we refer to this as the *depth structure of nearest neighbor*. Interpreting this together with Table I, the regions to consider are those shown in the figure: c/c, c/s, s/c, s/s, and regions with one side fixed.

Here, we interpret the meaning of vertical axis direction and change value. From (1), d_{NN_2} is the sum of the degrees of all words connected to W_j. This means that equation (1), by taking the sum of d_{j_α} for connected words, emphasizes the properties of the group of words that W_j can combine with.

When d_{NN_2} has a large value, W_j indicates that it has many links with global words, meaning it is a more general-purpose word. Concepts explained by using many general-purpose words generally correspond to explanations using examples or analogies, or descriptions based on deeper understanding by linking to other concepts. In Figure 4, a strong peak in the c/c region indicates the presence of careful explanations using examples. A strong peak in the s/c region signifies that the specific word itself represents a concept capable of having many connections, and that it is being explained more concretely using technical terms and examples. In either case, it indicates the presence of descriptions that utilize newly acquired knowledge.

When d_{NN_2} has a small value, W_j has only links to specific words. This means that even if W_j were a global word, it would only be used to explain specific words. That is, it indicates that after learning the concept of a new item, learners attempted to explain it using the exact learned phrase without much elaboration. This tendency is commonly observed immediately after learning the concept of a new item. Particularly when a strong peak appears in the s/s area in Figure 4, it is reasonable to assume that many learners absorbed the new item but remain in a state of mere absorption.

Considering the above points comprehensively, the distribution of $(d_{NN_1},\ d_{NN_2})$ enables estimation of temporal transitions when administering similar tasks after a period following training. Although a strong peak is observed in the s/s region immediately after learning, as learning progresses, it changes in the vertical or horizontal direction. If it transitions to c/c via s/c or c/s, it indicate that standard learning deepening has been achieved.

In the case of analysis toward actual short answer data, we can easy image that such distributions do not necessarily appear in a single specific location. Therefore, we consider synthesizing K probability density functions and estimating learners' acquisition of new concepts by determining where the composite result shows strong responses.

B. The Effect of the Standard Deviation in p_k on the Mixture Model

Figure 5 shows the computation results for the influence of $\sigma_{NN_1},\ \sigma_{NN_2}$ for each p_k on the mixture model distribution. Here, $\mu_{NN_1},\ \mu_{NN_2}$ for p_1 were set to $(3,\ 3)$, $\mu_{NN_1},\ \mu_{NN_2}$ for p_2 as $(3, 1)$, and ρ fixed at 0.5. $\sigma_{NN_1},\ \sigma_{NN_2}$ were varied from 1 to 3. In the figure, the top row shows the mixture model distribution when $\sigma_{NN_1} = \sigma_{NN_2}$, the middle row shows the distribution when σ_{NN_1} for p_1 is set to 1 and the other standard deviations are varied, the bottom row shows the distribution when σ_{NN_1} for p_1 is set to 2 and the other standard deviations are varied.

The upper right section shows that both p_1 and p_2 have $(\sigma_{NN_1},\ \sigma_{NN_2}) = (3,\ 3)$. The distribution is a diffuse witout specific concentration large value. Therefore, we concluded that creating a template is unnecessary when the standard deviation is excessively large, and thus created combinations with standard deviations of either 1 or 2.

Overall, when p_k includes $(\sigma_{NN_1},\ \sigma_{NN_2}) = (1,\ 1)$, there is one location showing a very strong peak, but elsewhere the distribution appears diffuse. Furthermore, the shape during diffusion can be reproduced as either a distorted shape or a clean elliptical shape depending on the combination of standard deviations within p_k.

C. The Effect of the Correlation Coefficient in p_k

Figure 6 investigates the influence of ρ_1 and ρ_2 for each p_k on the distribution of the mixture model. Here, $\mu_{NN_1},\ \mu_{NN_2}$

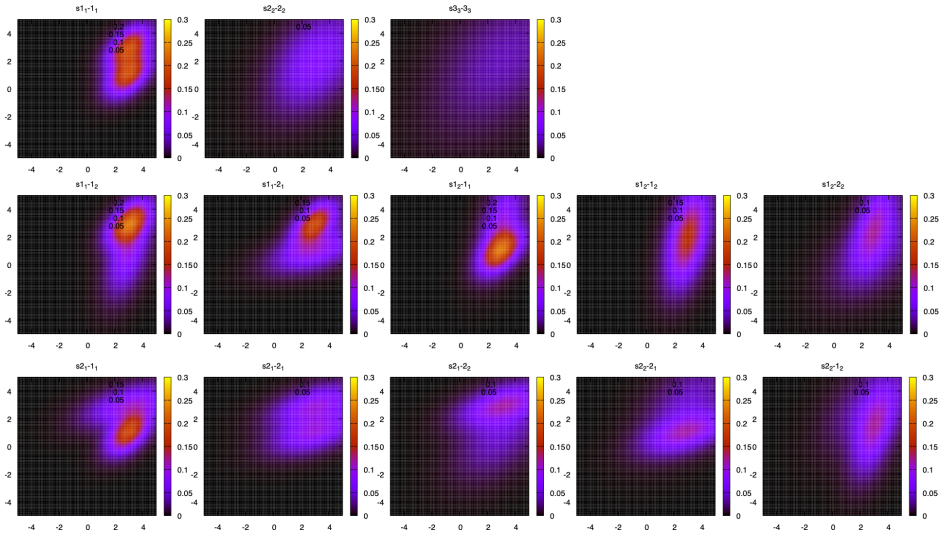


Figure 5. Two dimensional normal distribution for $K = 2$ when fixed at $(\mu_x,\ \mu_y) = (3,\ 3), (3,\ 1), \rho = 0.5$. The figure does not show cases where either σ_{NN_1} or σ_{NN_2} equals 3.

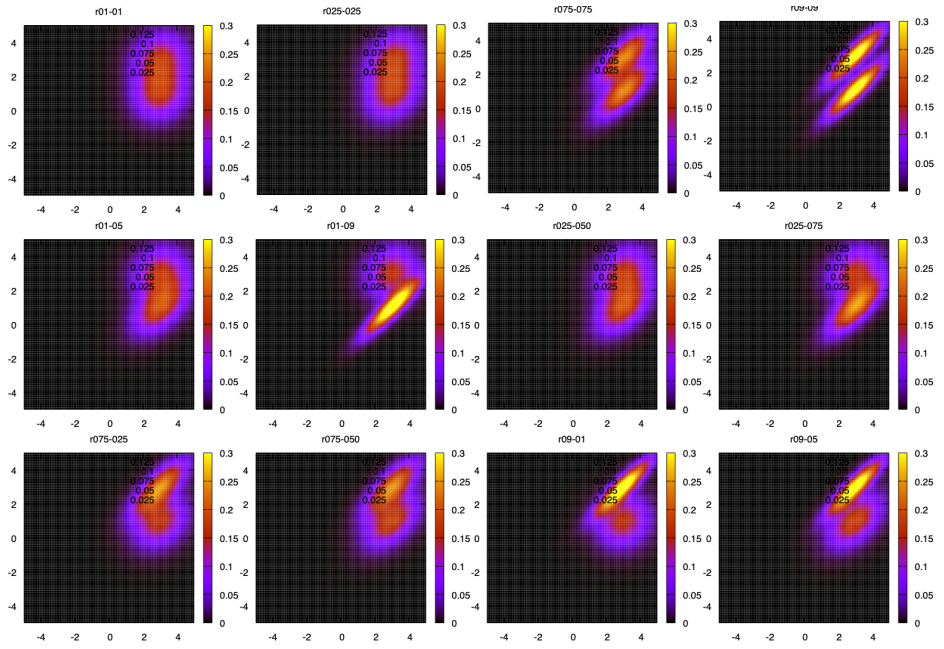


Figure 6. Two dimensional normal distribution with $K = 2$ when $\mu_x, \mu_y = (3, 3)$, $\mu_x, \mu_y = (3, 1)$, and $\sigma_x, \sigma_y = (1, 1)$ are fixed.

for p_1 were set to $(3,\ 3)$, $\mu_{NN_1},\ \mu_{NN_2}$ for p_2 as $(3, 1)$, and $\sigma_{NN_1},\ \sigma_{NN_2}$ were varied from 1 to 3. In the figure, the top row shows the $\rho_1 = \rho_2$, the middle row shows the $\rho_1 < \rho_2$, the bottom row shows the $\rho_1 > \rho_2$.

As a general trend, as the value of ρ increases, p_k shows a concentration of spread in specific directions toward the center, and the probability density takes on sharply large values. That is, it exhibits bias. However, this bias appears to occur generally around $\rho > 0.7$. Regarding shape, when the centers of p_k are close together, the diffusion tendency is not lost even if the values of ρ for each k are small. Directivity becomes pronounced only when large values of ρ are observed on one side exclusively.

D. The Effect of Average and Correlation Coefficient on the Mixture Model in p_k

The figure depicts a mixture model of two dimensional normal distributions, with σ_{NN_1} and σ_{NN_2} each fixed at 1 for p_k. In the figure, The top row fixes $(\mu_{NN_1},\ \mu_{NN_2})$ at $(0.5,\ 3)$, $(0.5,\ -3)$ The middle row fixes $(\mu_{NN_1},\ \mu_{NN_2})$ at $(1,\ 1)$, $(-1,\ -1)$. The bottom row fixes $(\mu_{NN_1},\ \mu_{NN_2})$ at $(3,\ 3)$, $(3,\ 1)$. Additionally, from the left column to the right column, the values of $(\rho_{p_1},\ \rho_{p_2})$ were varied to $(0.1,\ 0.1)$, $(0.1,\ 0.5)$, $(0.5,\ 0.1)$, $(0.5,\ 0.5)$. The displayed colors represent probability density. As the color transitions from blue to yellow, it indicates a higher probability density for the corresponding $(d_{NN_1},\ d_{NN_2})$, meaning the frequency

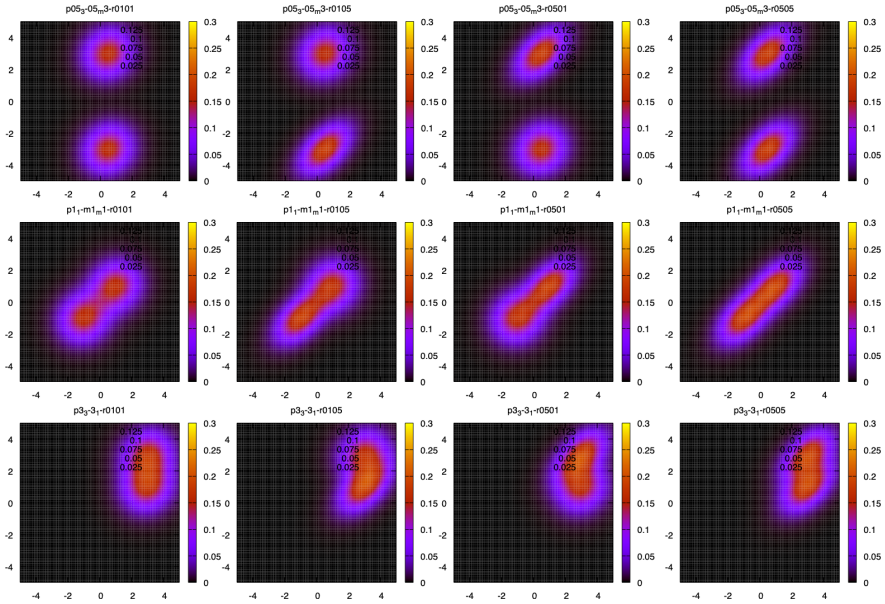


Figure 7. Two dimensional normal distribution for $K = 2$ when fixed at $(\sigma_x,\ \sigma_y) = (1,\ 1)$.

of occurrence for the degree with that value increases.

Comparing the upper and lower parts of the figure, when $\mu_{NN_1},\ \mu_{NN_2}$ are sufficiently far apart, $p_1,\ p_2$ have independent distributions. However, as $p_1,\ p_2$ approach each other, the two distributions are mixture. Specifically, when varying $(\rho_{p_1},\ \rho_{p_2})$ up to $(0.5, 0.5)$, depending on the relative positions of $(\mu_{NN_1},\ \mu_{NN_2})$ between p_k, they may appear to form a single elliptical distribution. Furthermore, comparing the left and right sides, as the values of ρ_{p_1} and ρ_{p_2} increase—that is, as the correlation strengthens—the distributions of both extend along the major axis of the ellipse, distorting their shape. However, no single point exhibits a decisively high density; instead, relatively widespread areas with moderately high values are observed.

E. The meanning of k, μ, σ, ρ in the Depth Structure of Nearest Neighbor

Based on the above, we now discuss the meaning of k, μ, σ, ρ in the proposed depth structure of nearest neighbor.

First, k indicates the number of core probability distributions in the depth structure of the nearest neighbor plane. When all respondents provide the same answer, the expected k would be 1 or a very small value. As responses become more diverse, the expected value of k increases. However, if there is too little consensus among respondents, setting k becomes meaningless, and the probability distribution for the plane will essentially diverge.

$(\mu_{NN_1},\ \mu_{NN_2})$ indicates positions on the same plane. If $(\mu_{NN_1},\ \mu_{NN_2})$ are separated between p_k, we consider the interference between p_k is able to be negligible. Since $(\mu_{NN_1},\ \mu_{NN_2})$ for p_k indicates the properties of connecting words for W_j, the existence of isolated high probability density regions at specific positions implies that the respondent group is using words with similar characteristics. Particularly when the number of k is small, we consider that learners' preferences lack diversity due to specific social factors like lectures or habits. When focusing on short answers to semi-open-ended questions, responses are mostly based on learned concepts, so the words contained in short answers should be somewhat limited. In this sense, it is unlikely that k would be large.

Next, $(\sigma_{NN_1},\ \sigma_{NN_2})$ indicates the degree of diversity within the same plane. The magnitude of σ_{NN_1} for a given p_k is expected to be small, as using uniform words for explanatory concepts likely restricts the number of connectable words. Conversely, the trend for σ_{NN_2} will vary significantly depending on how extensively explanations are provided or whether additional conceptual explanations are given. If additional explanations of the concept are provided before a deep understanding of the learned content is achieved, we consider the σ_{NN_2} will be tend to small. Conversely, when providing additional explanations of the concept, the diversity of words used exists depending on the perspective from which the additional explanation is made. Therefore, we consider the σ_{NN_2} will be tend to large.

Finally, we consider the meaning of ρ_k in the plane. The ordinal sense of a correlation coefficient indicates the strength of the relationship between d_{NN_1} and d_{NN_2}. While d_{NN_1} and d_{NN_2} cannot be expressed as functions describing this relationship, they can be positioned as properties of W_{j_α} that are connectable to W_j. Since most W_j are considered linkable to both global words and specific words, ρ is expected to take very small values. However, for p_k regarding specific words, if only words used exclusively for explaining specialized concepts can be employed, then ρ would likely exhibit strong correlation. In this sense, ρ should be useful for exploring the presence or absence of such special conditions.

From the above, it indicates that when analyzing p_k on the depth structure of the nearest neighbor plane, μ and σ are the

most important factors, but occasionally ρ is also necessary. Moving forward, it will be necessary to demonstrate the validity of this series of considerations regarding the parameters when applying them to actual data.

IV. CONCLUSION AND FUTURE WORKS

In this paper, we proposed a framework that classifies learners' short answers into three levels as one method for grading semi-open-ended questions, requiring only minimal knowledge of complex networks. This framework enables the classification of learners' level of understanding and degree of knowledge expression, and contributing to subsequent educational efforts. By adapting the concept of the mixed Gaussian model, we created a depth structure of the nearest neighbor plane for WCN. We demonstrated that by adjusting the number of p_k distributed within it and the effects of variations in μ, σ, and ρ, as well as the synthesis process, it is possible to reproduce various distributions. In practice, obtaining approximations of probability density for this plane it requires preparing real data and deriving approximate probability density using methods like the EM algorithm. Regarding real data, we plan to validate this approach using short answer data from thousands of semi-open-ended questions on graph reading comprehension, for example, obtained by Yagashira et al. [19]. Additionally, a key advantage of this method is its independence from the language used. We aim to apply it to other languages in the future to confirm its usefulness. We believe this method is beneficial when semi-open-ended questions can be posed, assuming there are settings like education or lectures where learners share a common understanding. On the other hand, considering the contexts in which this approach is applied, we assume that the realistic application of this method is for small-scale settings, such as within a classroom or a school. For this reason, while application to large-scale datasets is not currently anticipated. If we consider applying this approach to large-scale datasets in the future, instability—such as an excessive number of degree —may arise, requiring appropriate measures to address it.

REFERENCES

[1] M. Sedighi, "Using of co-word analysis method in mapping of the structure of scientific fields(case study: The field of informetrics)", *Iranian Journal of Information Processing Management*, vol. 30, pp. 373–396, Dec. 2015.

[2] T. Li, J. Bai, X. Yang, Q. Liu, and Y. Chen, "Co-occurrence network of high-frequency words in the bioinformatics literature: Structural characteristics and evolution", *Applied Sciences*, vol. 8, no. 10, 2018, ISSN: 2076-3417.

[3] W. Liang, "Spectra of english evolving word co-occurrence networks", *Physica A: Statistical Mechanics and its Applications*, vol. 468, pp. 802–808, 2017, ISSN: 0378-4371.

[4] M. Garg and M. Kumar, "The structure of word co-occurrence network for microblogs", *Physica A: Statistical Mechanics and its Applications*, vol. 512, pp. 698–720, 2018, ISSN: 0378-4371.

[5] M. Fudolig, T. Alshaabi, M. Arnold, C. Danforth, and P. Dodds, "Sentiment and structure in word co-occurrence networks on twitter", *Applied Network Science*, vol. 7, p. 9, Feb. 2022.

[6] D. Amancio, J. Machicao, and L. Quispe, "Leveraging word embeddings to enhance co-occurrence networks: A statistical analysis", *PLOS One*, vol. 20, Jul. 2025.

[7] T. Millington and S. Luz, "Analysis and classification of word co-occurrence networks from alzheimer's patients and controls", *Frontiers in Computer Science*, vol. 3-2021, 2021, ISSN: 2624-9898.

[8] K. Zupanc and Z. Bosnić, "Increasing accuracy of automated essay grading by grouping similar graders", in *Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics*, ser. WIMS '18, Novi Sad, Serbia: Association for Computing Machinery, 2018, pp. 1–6, ISBN: 9781450354899.

[9] T. Wang, N. Inoue, H. Ouchi, T. Mizumoto, and K. Inui, "Inject rubrics into short answer grading system", in *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, C. Cherry et al., Eds., Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 175–182.

[10] A. Condor, Z. Pardos, and M. Linn, "Representing scoring rubrics as graphs for automatic short answer grading", in *Artificial Intelligence in Education*, M. M. Rodrigo, N. Matsuda, A. I. Cristea, and V. Dimitrova, Eds., Cham: Springer International Publishing, 2022, pp. 354–365, ISBN: 978-3-031-11644-5.

[11] L. Zhang, Y. Huang, X. Yang, S. Yu, and F. Zhuang, "An automatic short-answer grading model for semi-open-ended questions", *Interactive Learning Environments*, vol. 30, no. 1, pp. 177–190, 2022.

[12] T. Lertchaturaporn and P. Songmuang, "Automated scoring model for semi-open-ended question in scientific explanation", in *2024 IEEE International Conference on Cybernetics and Innovations (ICCI)*, 2024, pp. 1–5. DOI: 10.1109/ICCI60780.2024.10532484.

[13] R. Conijn, P. Kahr, and C. Snijders, "The effects of explanations in automated essay scoring systems on student trust and motivation", *Journal of Learning Analytics*, vol. 10, no. 1, pp. 37–53, Mar. 2023.

[14] F. Jamil and I. A. Hameed, "Toward intelligent open-ended questions evaluation based on predictive optimization", *Expert Systems with Applications*, vol. 231, p. 120 640, 2023, ISSN: 0957-4174.

[15] L.-H. Chang and F. Ginter, "Automatic short answer grading for finnish with chatgpt", *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 21, pp. 23 173–23 181, Mar. 2024.

[16] C. Grévisse, "Llm-based automatic short answer grading in undergraduate medical education", *BMC Medical Education*, vol. 24, Sep. 2024.

[17] R. Ferreira Mello et al., "Automatic short answer grading in the llm era: Does gpt-4 with prompt engineering beat traditional models?", in *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, ser. LAK '25, Association for Computing Machinery, 2025, pp. 93–103, ISBN: 9798400707018.

[18] K. T. Nakahira, M. Watanabe, and M. Kitajima, "Assessment of developmental stages of generic skills: A case study", in *Proceeding of the 22nd International Conference on Computers in Education, ICCE 2014, Nara, Japan, November 30 - December 4, 2024*, Asia-Pacific Society for Computers in Education, 2014, pp. 200–205.

[19] Y. Yagashira, K. T. Nakahira, M. Arai, G. Kumoi, and T. Yukawa, "Proposal of a method for estimating the acquisition of graph comprehension ability based on students' descriptive answers", *Procedia Computer Science*, vol. 270, pp. 1856–1865, 2025, 29th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2025), ISSN: 1877-0509.

AMICA: Accessible Multimodal Interaction Conversational Assistant for School Children with Intellectual Disabilities

André Frank Krause
Rhine-Waal University of Applied Sciences
Kamp Lintfort, Germany
e-mail: andrefrank.krause@hochschule-rhein-waal.de

Artem Savelov
Rhine-Waal University of Applied Sciences
Kamp Lintfort, Germany
e-mail: artem.savelov@hsrw.org

Carrie Ching
Rhine-Waal University of Applied Sciences
Kamp Lintfort, Germany
e-mail: kar-wai-carrie.ching@hsrw.org

Kyra Kannen
Rhine-Waal University of Applied Sciences
Kamp Lintfort, Germany
e-mail: kyra.kannen@hochschule-rhein-waal.de

Karola Pitsch
University Duisburg-Essen
Essen, Germany
e-mail: karola.pitsch@uni-due.de

Nele Wild-Wall, Christian Ressel
Rhine-Waal University of Applied Sciences
Kamp Lintfort, Germany
e-mail: {nele.wild-wall | christian.ressel}@hochschule-rhein-waal.de

***Abstract*—This paper reports on progress in the development of an Accessible Multimodal Interaction Conversational Assistant (AMICA) for children with intellectual disabilities. Early access to technologies based on artificial intelligence during childhood is essential to promote inclusivity and to reduce the digital divide. To protect the vulnerable user group, AMICA has a strong focus on privacy through exclusive use of open source technologies and locally executable artificial intelligence models. The voice assistant uses a three-stage system architecture for low-latency information retrieval in local, domain-specific databases. Common questions are answered with low latency in stage 1, where answers are retrieved from a curated question & answer database using semantic search. Context sensitive questions or more general queries not represented in the question & answer database will be escalated to stage two (context-based rephrasing), or further to stage three (Large Language Model-based answer generation). The user experience of the voice assistant was tested in two pilot studies. The first study observed the interaction of students with the system from a conversation analysis perspective. It revealed an issue with semantic search in stage one, if multiple questions occur in a single user query. The second study was performed at a school for children with intellectual disabilities. Main results: 1. Automatic speech recognition failed for children with speech disorders. 2. Large Language Model-generated answers are often too complex and need to be simplified into "easy language". 3. The current, strictly turn-based voice interaction using a Push-to-Talk microphone posed a challenge for some children. The studies substantiate the importance of inclusive design for accessible assistive technologies as well as the need for inclusive speech recognition and easy language generation.**

***Keywords*-multimodal dialogue systems; intellectual disabilities; inclusive design; data privacy; conversation analysis.**

I. Introduction

The United Nations' Convention on the Rights of Persons with Disabilities [1] demands that digital technologies, including Artificial Intelligence (AI), must be designed such that people with disabilities can access these tools to enable their effective participation and inclusion in society. This paper reports on progress in the development of an Accessible Multimodal Interaction Conversational Assistant (AMICA) for school children with intellectual disabilities.

AMICA aims to ease access to modern AI technologies and information retrieval with a strong focus on privacy. A fully privacy-respecting system is of high importance to protect the target user group, enabling their right of informational self-determination [2]. The current, cloud-dominated AI landscape often lacks clear AI regulations and privacy guarantees. According to the Artificial intelligence index report 2025 [3], below 50% of global users trust that AI companies protect their personal data. Therefore, AMICA exclusively uses open source technologies and open AI models that can be executed locally on consumer-grade hardware without an internet connection. This approach enables the best possible privacy, low deployment costs and resilience from internet disruptions. AMICA was implemented in strict accordance with our design principle of "100% privacy, 0% cloud".

The system allows easy access to domain-specific knowledge relevant to the school children, e.g., information about their school, teachers, room locations, schedules, the student-canteen menu and other relevant data. The data is stored as question-answer pairs in a table that can be easily edited and extended by the school staff.

We hope that AMICA lowers the inhibition threshold for asking certain questions that children might find embarrassing. Our assumption is that children with intellectual disabilities may hesitate less asking AMICA specific questions, compared to approaching classmates or teachers.

Large Language Models (LLMs) encode extensive factual knowledge, enabling a paradigm shift in information retrieval (e.g., for search engines, dialogue- and question-answering

 ISBN: 978-1-68558-375-0

systems [4]). Further, LLMs often show surprising, non-anticipated emergent properties compared to smaller AI-models [5]. Examples include high quality code generation and In-Context Learning (ICL). ICL is the ability of LLMs to execute novel tasks without expensive retraining [6] by prompting the model with just a handful of examples and instructions [7]. Therefore, LLMs can provide a foundation to implement modern interactive assistive technologies, for example, conversational agents.

Unfortunately, LLMs exhibit a number of problematic issues, the most prominent being hallucinations: The generation of seemingly plausible, yet factually incorrect or fabricated content [8][9]. LLMs present such hallucinated content in a convincing, human-like way [8][10] that is difficult to detect. Dilmegani and Daldal [11] recently tested prominent commercial LLMs and found hallucination rates ranging between 15% and 52%. Other issues include the emergence of harmful capabilities (e.g., deception, manipulation, reward hacking) [5], failures in deep reasoning [12] and the lack of coherent world models [13][14], leading to inconsistent, brittle performance even in related tasks with comparable complexity [13]. A highly undesired property of LLMs is that seemingly innocuous prompts may trigger unintended chatbot behaviors, like excessive sycophancy or toxicity, potentially causing AI-related psychological harm, as reviewed in [15].

An established method to mitigate hallucinations is Retrieval Augmented Generation (RAG) [16]. RAG retrieves relevant information from external knowledge bases and supplies the selected knowledge together with the user prompt to the LLM. RAG significantly enhances answer accuracy and provides domain-specific knowledge to the LLM [16].

AMICA employs a three-stage information retrieval- and answer-generation approach, as detailed in Section II-B. The first stage uses semantic search over a table with question-answer pairs, based on sentence embedding using a sentence transformer model. Sentence transformer models [17][18] analyze the contextual meaning of words in a sentence in both directions, capturing nuanced, deeper semantic meanings of sentences. A good semantic match to a user question results in direct answer output (see Figure 3, avoiding any of the aforementioned generative model issues. This three-stage approach has the intentional advantage of very short response times for stage one. Low-latency responses improve the quality of natural language interaction [19].

The remainder of the paper is organized as follows: In Section II, the technical details of the system architecture are described, including subsections about speech recognition, answer generation and speech synthesis. Section III presents the results of two pilot studies, followed by a discussion in Section IV.

II. System Architecture

AMICA uses a scalable, distributed architecture. It is composed of several modules (see Figure 1) that run in parallel and communicate asynchronously via message queues. These modules can run on different nodes within a local network or on the same machine. For example, the answer generation module may run on a separate machine with a sufficient amount of memory and compute to execute the AI models for answer generation.

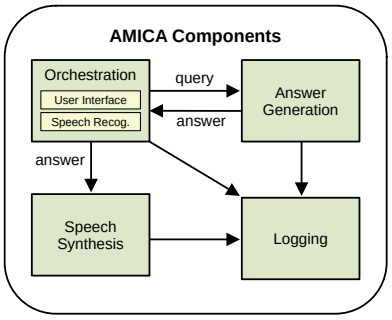


Figure 1. System Architecture. AMICA uses a modular, scalable and distributed architecture consisting of four modules: 1. an orchestration module including user interface and speech recognition sub-modules; 2. an answer generation module; 3. a speech synthesis module and 4. a module for optional logging. These modules can run on different compute nodes within a local network and communicate through message queues.

Figure 2. A person interacting with the voice assistant AMICA using a PTT microphone. After releasing the speak-button, the system generates an answer (see Figure 3) that is shown on a monitor and reads it aloud. Background Artwork: Georgina Chacón, "Mystical Llama" CC BY-NC-ND 3.0.

A. Speech Recognition

The primary interaction method with AMICA is voice input. Children who cannot speak may use a keyboard or a touch display with icons as an alternative input method, which is planned as a future feature. The voice input of a user is captured by a Push-to-Talk (PTT) microphone (Figure 2). A PTT-microphone offers a notable privacy advantage: As long as the talk-button is not pressed, the microphone is internally short-circuited and no information can accidentally enter the system. Further, pressing and holding the speak button clearly indicates user-intent to interact with the system, avoiding unreliable methods like keyword based system activation (e.g., "Hey AMICA!") or volume threshold-based start- & end-of-speech detection. The captured voice signal is processed using Whisper-large-v3-turbo-german [20], a Transformer based speech recognition model based on Whisper [21] and fine-tuned for German speech.

B. Answer Generation

The response to a users query is generated in up to three stages (see also Figure 3), as detailed in the following three subsections. In stage one - the retrieval phase - a question is

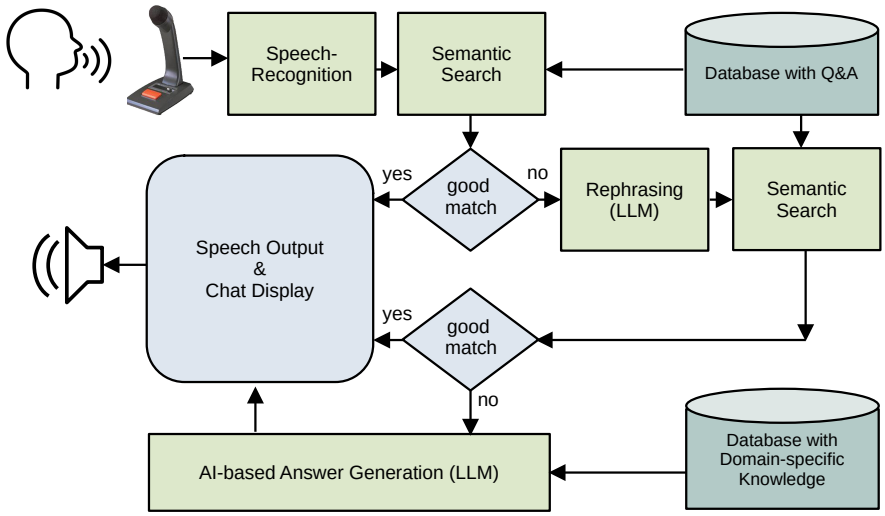


Figure 3. AMICA data-flow diagram. The response to a query is generated in up to three stages. Stage 1 - Retrieval: The question or query of the user is compared with all available questions in the Q&A database using semantic search. If the matching score is high enough, the corresponding answer is directly displayed and read aloud. Stage 2 - Rephrasing: The user query is combined with context information from the chat history, rephrased and scored. Stage 3 - Generation: If the score is still not high enough, a broader answer is provided using LLM-based retrieval augmented generation.

compared with all available questions in the Question & Answer (Q&A) database using semantic search. If the matching score is high enough, the corresponding answer is directly displayed and read aloud with low latency. In stage two, the user query is combined with context information from the chat history, rephrased and scored. If the score is still not high enough, stage three generates a broader answer using LLM-based retrieval augmented generation. At stages one and two, there is no risk of LLM hallucinations because answers are always factual and are output exactly as stored in the Q&A database.

1) Retrieval: At the first stage, the system tries to retrieve a suitable answer to a user's query by finding the semantically closest information in the Q&A database. The retrieval is performed using semantic search: The query is transformed into a numerical vector using an embedding model. This vector is compared with the vector representations of the questions in the Q&A database using a similarity metric (cosine-similarity). A high similarity score indicates a high semantic overlap. If this score exceeds a predefined threshold, the best matching answer is directly presented to the user (see Section II-C). The embeddings are generated using Arctic Embed 2.0 L, a multi-lingual, enterprise grade, open-weight model [22].

2) Rephrasing: If a follow-up question or query implicitly relies on previous context, direct retrieval (stage one) may result in a low semantic matching score. In such a situation, the question can be "rephrased" based on the previous chat history with the user. Rephrasing adds context information that might be necessary to retrieve an answer from the Q&A database.

Here is a sample dialogue:

- User: How can i go home?
- AMICA: You can take the bus.
- User: When is it departing?
- AMICA data-flow:

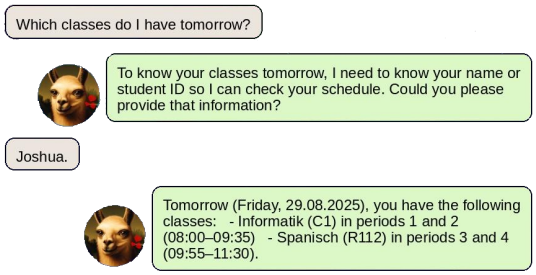


Figure 4. The assistant displays the conversation thread in an established layout similar to the chat history of typical messenger apps.

1) Retrieval: low score
2) Rephrasing: When is my bus departing?
3) Retrieval: The bus departs at 4:30 p.m.

A LLM is instructed to perform the rephrasing task by a specific rephrasing prompt. This prompt includes the last three messages from the chat history of the current session.

3) Augmented Generation: If the rephrasing step did not result in a good semantic match with any Q&A database entry, the rephrased query will be answered directly by a LLM. The answer generation uses the "History Aware Retriever" [23] method: The LLM generates the answer based on the last 10 messages from the chat history with an addition of relevant information from a database with domain-specific knowledge. Relevant information is retrieved by semantically matching the rephrased query to chunks of information in this database.

The domain-specific database (see Figure 3) contains knowledge about the intended application area, currently a school for children with intellectual disabilities. In contrast to the Q&A database, the knowledge in this database can be enumerations, free-form texts, data scraped from websites (e.g., the current weather) and short facts.

Table 1: conversation analysis	
(a)	Participants' two-part question: Two participants – PAR01 and PAR02 – ask a two-part question to AMICA asking "and when is lunchtime and where do we get lunch," (01: PAR02).
(b)	AMICA's first reply – first part only: In its reply, AMICA only answers the first part "lunch begins at twelve o'clock zero zero o'clock" (l. 04).
(c)	Participants' repeated question – second part: PAR01 treats AMICA's answer as incomplete inquiring again for the second part "and where do we get lunch?" (l. 05).
(d)	AMICA's second reply – first part only: Again, AMICA's answer only addresses the first part of the initial question (although only the second part had been asked for) by replying "lunch begins at twelve o'clock zero zero o'clock" (l. 10). The participants finally resign and address a new question to the system (l. 11).

Table 2: conversation analysis, including log files	
(b*)	AMICA's first reply – "semantic search": The log files reveal that PAR02's initial two-fold question (l. 01) has been correctly received by the "speech-to-text" component as "and when is lunchtime and where do we get lunch," (l. 02). In the system's next step, the "semantic search" component, however, reduces the two-fold question only to the first part (i) "When is lunchtime?" (l. 03), and thus produces the answer accordingly (l. 04).
(d*)	AMICA's second reply – "semantic search": The same phenomenon occurs during the participants' repeated question asking only for second part of the initial two-fold question: "and when do we get lunch" (l. 05). Again, the "speech-to-text" component receives PAR01's input correctly (l. 06), whereas the "semantic search" component reformulates it to the first part of the question (which has not been uttered at this time): "When is lunchtime?" (l. 07).

C. Speech Output and Chat Display

Piper, a fast and local neural text-to-speech engine [24], is used to synthesize friendly, German-language speech. The module uses an open weight model from a collection of piper-compatible models [25] trained on the Thorsten-Voice dataset. Thorsten-Voice is an open source (CC0 license) German voice dataset containing 40 hours of transcribed voice recordings [26].

The chat history is printed on-screen using an established and commonly recognized layout familiar from typical instant messaging applications (Figure 4).

D. Implementation Details

The primary development and execution platform was Linux Mint 22.2. AMICA was implemented using the Python programming language version 3.12. Process-based parallel execution of the different modules is facilitated using Python's multiprocessing package. The Graphical User Interface (GUI) currently uses OpenCV for window management and interface elements rendering. It is rendered using an immediate mode GUI approach [27] without any further dependencies. True-type fonts are rendered using the Python Imaging Library (PIL). LLMs are provided by Ollama [28], a tool for management and local execution of open weight models, and accessed using Ollama's REST API. Embedding generation, matching and storage (in-memory vector database) are provided by the LangChain package. Audio from the PTT-microphone is recorded by the PyAudio module and preprocessed with the Librosa package. The PTT button press is registered by a Raspberry Pi Pico 2, programmed and configured with the Belay Python library.

III. Pilot Studies

The voice assistant was tested in two pilot studies. Specific user experience issues were found that need to be taken into account in the further development of the voice assistant.

A. Pilot Study 1

Pilot study 1 aimed at understanding how users would attempt to communicate with AMICA and at evaluating how the system's architecture and dialogue features might support the human-machine interaction. Pilot study 1 was devised as a semi-experimental setup in which pairs of users were asked to assume the role of German-speaking students who arrived at a new school and used AMICA to find information about routines and spare time activities of their new school.

The pilot study was conducted in May 2025 with 14 voluntary students from the Faculty of Humanities (German proficiency level: native speakers) of a Germany university, resulting in seven trials. Each trial lasted between 13 and 20 minutes (total duration: 126 minutes). The sessions were recorded with two external video cameras, while the system's actions, states and decisions were logged internally. The log files comprise of internal and external information from the different modules with the following tags: (a) speech_input (stored as .wav files), (b) speech-to-text, (c) semantic_search, (d) history_aware_retriever, (e) sentence_generation, and (f) speech_output (stored as .wav files).

Video recordings and log files were synchronized manually and imported into the timeline-based transcript editor ELAN [29] so that the conversation analysis can relate external observation and internal information. The analysis was adapted from Ethnomethodological Conversation Analysis to investigate human-machine interaction (e.g., [30]).

Initial exploration of the data shows that participants were able to use AMICA intuitively to obtain information and to plan some spare-time activities. However, inspecting the user interactions with AMICA in greater details, we found a set of instances in which the system's responses leave room for optimization.

Table 1 shows a detailed analysis of the conversation fragment of trial 01, 01:23 – 01:50, as shown in Figure 5. The transcription follows the GAT-convention. The interaction with AMICA was conducted in German language. English translation was added in bold below each line of German text in Figure 5. The analysis revealed that a two-part question was not properly handled by the system. Attempting to gain a better understanding of how this problem emerged, we included, in a second analytical step, the system's log files into the analysis (Table 2).

Pilot study 1 revealed that compound queries (i.e., queries with multiple questions) are not properly handled yet in the

first stage of AMICA's system architecture. Several options may solve the issue:

1) A trivial - but not particularly user-friendly solution - is to instruct users to ask only one question per query.
2) Queries may be split into individual questions that are answered sequentially by the first stage.
3) The semantic-matching threshold may be fine-tuned to avoid a false match due to a relatively low threshold.
4) More sophisticated semantic search strategies, e.g., using a dynamic threshold and considering the top-k matches, should be explored.
5) New sentence transformer models should be tested regularly to determine whether they can improve the system's semantic search performance.
6) If a question is repeated, the user indicates dissatisfaction with the given answer. In such situation, the system should jump directly from the first stage to LLM-based answer generation (the third stage), where the LLM will be able to answer compound queries.

B. Pilot Study 2

An early-stage exploratory pilot test was conducted with N = three children with intellectual disabilities (two males and one female, aged from 10 to 14 years) to uncover design challenges and initial usability and user experience issues. The primary objective of the study was to observe and analyse interaction patterns with the conversational assistant. Only children who were able to speak and hear as well as had no visual impairment were included in the pilot study. From those children whose parents gave informed consent, their class teacher randomly selected three participants. The participation was voluntary, and the children could end the experiment early at any time. The experiment was conducted in a quiet, small meeting room of their school. In each pilot test session, a participating child engaged with the system individually for approximately 10 minutes, under the supervision of a familiar reference person (their class teacher). All three sessions were conducted on the same day.

We have adhered to the applicable ethical principles in the 1964 Declaration of Helsinki [31] such that the health and well-being of the participants were considered. Ethical issues have been managed appropriately:

- Data privacy: No audio or video recordings were made. No personally identifiable information was requested. As mentioned above, the system operates offline, disconnected from the internet. Hence, there is minimal risk of data breaches.
- Informed consent: The consent form was generated using an electronic tool for the compilation of informed consent documents, which was developed by the Ethics Committee of the Technical University of Munich [32].
- Risks: It is possible that the children may confuse the system as a real human. In such case, the class teacher would explain to the child that it is an artificial system. However, it was not needed in the three sessions.

Observations and feedback from the teacher revealed the children's initial hesitation, strong needs for validation, and uncertainty about the capabilities of the system. For instance, participants frequently sought confirmation from the teacher before starting to interact with the system and were unsure about what questions they could ask. However, after a warm-up phase, the children started interacting with the system on their own. For two of the three participants, the press-to-speak mechanism and strict turn-taking interaction posed a challenge. Frequently, participants attempted to speak while the system was still responding, indicating a mismatch between system design and natural conversational behaviour. It appeared that participants might benefit from additional clarification and confirmation provided by the system, e.g., by validating or rephrasing their questions. One of the most critical findings of this pilot study was the system's difficulty in understanding children with articulation disorders, which frequently led to breakdowns in interaction and prevented successful assistance.

IV. Discussion

The second pilot study revealed important usability and interaction challenges that need to be considered in the next design iteration.

First, initial onboarding and real-time guidance should be provided by the system in order to reduce uncertainty and support independent use. The integration of confirmation and clarification strategies, in conjunction with an interactive design that fosters autonomy, is imperative to enhance confidence and promote independent engagement with the system.

Second, flexible turn-taking mechanisms and equally robust alternatives to Press-To-Speak input are needed to better accommodate natural communication behavior and diverse motor or cognitive abilities. AMICA currently uses a conventional but flexible cascading approach that combines different modules for speech recognition, text generation, and speech synthesis into a pipeline. This modular approach is not without drawbacks [33]: A) Information loss: Certain paralinguistic cues are lost during the speech-to-text transformation. These cues include prosody-based emotional states, sarcasm, irony and other features not encoded in grammar and vocabulary. B) Error propagation: Inaccuracies in speech recognition will propagate down the pipeline, confusing semantic search and generative models in the text generation module. C) Latency: Processing delays add up due to sequential processing. Recent approaches (for review, see [33][34]) combine all modules into a single model that can process and generate spoken language, enabling a more natural end-to-end speech interaction. Such full-duplex, "Speech-to-Speech" language models can imitate natural human conversation patterns more closely, e.g., by simulating active listening, graceful handling of interruptions and simultaneous speaking of both the model and the user [34]. In addition to natural conversation flow, some open source models also address empathetic interaction [35].

Third, Automatic Speech Recognition (ASR) for children is challenging due to larger variations in children's speech compared to adult speech [36], it becomes even harder for

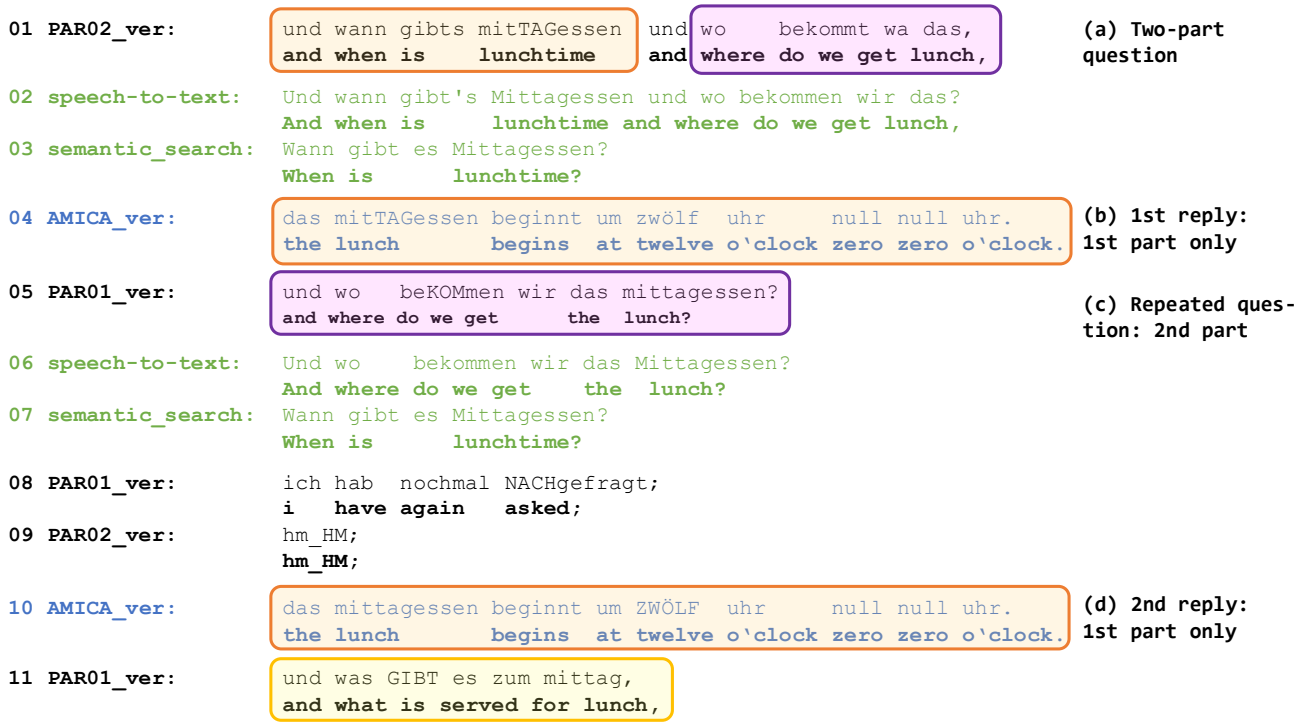


Figure 5. Transcript of verbal actions and log files of one interactional sequence in which two participants (PAR01 and PAR02) use the voice assistant AMICA in a semi-experimental setup.

children at special education schools as they may exhibit various forms of speech disorders. This highlights the necessity for research in the field of inclusive speech recognition. Personalizing ASR models using individual speech samples can improve speech recognition accuracy [37] and even outperform human listeners [38].

For example, [39] demonstrated that an existing machine-learning-based ASR model (Whisper) could be fine-tuned on speech samples of a German-speaking child with congenital speech disorders. The speech recognition accuracy improved for read speech, but not for conversational speech. There exist initiatives that aim to collect large and diverse corpora of speech samples from individuals with varied accents, dialects, or speech disorders as a foundation for ASR model training [40][41].

Fourth, the LLM-generated answers use a language style that often occurs to be too complex to grasp for the intended target group of children with cognitive disabilities. It is planned to employ some specialized LLM that directly generate "plain language", or a further simplified form of German language called "leichte Sprache" (literally: easy language) which was specifically designed for inclusivity. Such LLMs are fine-tuned on corpora of plain / easy language, as demonstrated for German text in [42].

Fifth, AMICA is designed as a standalone system, disconnected from the internet and local networks, to guarantee the best possible privacy and security. The lack of connectivity slightly diminishes the overall user experience because the databases cannot be updated over a local network but only by direct data transfer via a USB-stick ("sneaker-net"). To streamline the update of the internal databases, it is planned to connect AMICA to a local network using a low-cost data-diode that provides strong privacy guarantees by enforcing unidirectional data-flow. Using such a data-diode, information can flow into the system, but cannot escape or be exfiltrated (for details, see [43]).

V. Conclusion and Future Work

A three-stage architecture for an accessible, low-cost, low-latency and privacy-focused voice assistant was presented. This architecture can be applied to voice assistants in other application areas where data protection and confidentiality are paramount. Depending on the associated risks of generating inaccurate responses, organizations may choose to opt out of LLM-based response generation (stage three) and rely only on semantic search.

The system was evaluated in two pilot studies, revealing key usability issues for the target group, e.g., the need for easy language generation and inclusive speech recognition. Inclusive speech recognition may be necessary for many other user groups, such as foreigners who may speak German with an accent and grammatical errors, older adults who may speak more slowly and less fluently, people with cognitive impairments or physical disabilities, and adults with speech disorders for other reasons. Easy language generation may also be preferred by many users to reduce their cognitive loads, especially in a future society that may be dominated by Voice User Interfaces.

Building on iterative user-centered design testing, our next step is a comprehensive user study evaluating usability and user experience of an improved version of AMICA. This approach directly responds to the gaps identified in the literature, where voice assistants for people with disabilities or special needs are predominantly assessed for clinical effectiveness with limited attention to system interaction quality, usability, and user experience, and where heterogeneous methods hinder comparability [44]. By applying systematic, validated usability and user experience measures, our study aims to contribute evidence toward more standardized evaluation practices and safer, more trustworthy voice assistant deployment for vulnerable users.

In conclusion, an early and iterative, user-centred system design is necessary for the development of child- and disability-sensitive assistive chat bots.

A. Future Work

As discussed in Section IV, it is critical to integrate features that are specifically designed for inclusivity. It is planned to replace the current AI models with specialized LLMs that can generate language suitable for different skill levels. A further goal is automatic personalization to the individual needs of school children, e.g., automatic adaptation to specific speech disorders and language skills. Alternative input methods for non-verbal children will be integrated (e.g., icons on a touch screen).

The user experience of updating the knowledge databases will be improved, maintaining perfect privacy using a data-diode. In general, guidelines for inclusive design [45] will be applied to improve the target group's user-experience. It is planned to extend the scope of AMICA by performing an in-depth "People, Activities, Context, and Technology" (PACT) analysis [46]. The PACT analysis may include different target groups, e.g., elderly people with dementia. In this context, AMICA may be modified to support biographical cognitive stimulation that can improve mood, cognitive function and quality of life for people with dementia [47].

ACKNOWLEDGMENTS

This work was supported by the Ministry of Culture and Science of the State of North Rhine-Westphalia as part of the project "Center for Assistive Technology Rhine-Ruhr" (ZAT) (11/2023 to 10/2026, Grant No. PB22-076A and PB22-076D).

We thank Aaron Schneider and Seetu Shrestha from the Rhine-Waal University of Applied Sciences for their support in conducting pilot study 2. We thank Felix Bergmann, Anne Ferger and Thomas Schmidt from the University of Duisburg-Essen for supporting the data management of study 1.

Source Code Availability and Licenses. The source code of AMICA is released under the GNU General Public License, version 3 (GPL-3.0) and is available at https://github.com/afkrause/amica. We also thank Georgina Chacón for granting permission to use her artwork "Mystical Llama" (CC BY-NC-ND 3.0 License) as the background image of AMICA's GUI.

Author Contributions. AFK, AS, CC, KK and KP contributed to the manuscript. AFK, AS and CC developed components of AMICA. AS was the core AI developer. KP supervised pilot study 1 and carried out data analysis. CR supervises the sub-project AMICA. NW, CR and KP acquired funding. NW and CR coordinate the ZAT project.

REFERENCES

[1] United Nations, *Convention on the rights of persons with disabilities*, United Nations, Treaty Series, vol. 2515, p. 3, Adopted by General Assembly resolution A/RES/61/106 on 13 December 2006; entered into force 3 May 2008, 2006.

[2] Council of Europe, *European convention on human rights, article 8(1)*, https://www.echr.coe.int/documents/convention_ENG.pdf, retrieved: March, 2026, 1950.

[3] N. Maslej et al., "Artificial intelligence index report 2025", *arXiv preprint arXiv:2504.07139*, 2025.

[4] Y. Zhu et al., "Large language models for information retrieval: A survey", *ACM Transactions on Information Systems*, vol. 44, no. 1, pp. 1–54, 2025.

[5] L. Berti, F. Giorgi, and G. Kasneci, "Emergent abilities in large language models: A survey", *arXiv preprint arXiv:2503.05788*, 2025.

[6] B. Dherin, M. Munn, H. Mazzawi, M. Wunder, and J. Gonzalvo, "Learning without training: The implicit dynamics of in-context learning", *arXiv preprint arXiv:2507.16003*, 2025.

[7] Q. Dong et al., "A survey on in-context learning", in *Proceedings of the 2024 conference on empirical methods in natural language processing*, 2024, pp. 1107–1128.

[8] Y. Zhang et al., "Siren's song in the ai ocean: A survey on hallucination in large language models", *Computational Linguistics*, pp. 1–46, 2025.

[9] M. Cossio, "A comprehensive taxonomy of hallucinations in large language models", *arXiv preprint arXiv:2508.01781*, 2025.

[10] L. Huang et al., "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions", *ACM Transactions on Information Systems*, vol. 43, no. 2, pp. 1–55, 2025.

[11] C. Dilmegani and A. Daldal, *AI Hallucination: Compare top LLMs like GPT-5.2 in 2026*, retrieved: March, 2026, Dec. 2025.

[12] P. Shojaee et al., "The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity", *arXiv preprint arXiv:2506.06941*, 2025.

[13] K. Vafa, J. Y. Chen, A. Rambachan, J. Kleinberg, and S. Mullainathan, "Evaluating the world model implicit in a generative model", *Advances in Neural Information Processing Systems*, vol. 37, pp. 26 941–26 975, 2024.

[14] C. Robertson and P. Wolff, "Llm world models are mental: Output layer evidence of brittle world model use in llm mechanical reasoning", *arXiv preprint arXiv:2507.15521*, 2025.

[15] S. Karny, A. Baez, and P. Pataranutaporn, *Neural transparency: Mechanistic interpretability interfaces for anticipating model behaviors for personalized ai*, 2025. arXiv: 2511 . 00230 [cs.HC].

[16] Y. Gao et al., "Retrieval-augmented generation for large language models: A survey", *arXiv preprint arXiv:2312.10997*, vol. 2, no. 1, 2023.

[17] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks", in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019.

[18] L. Stankevičius and M. Lukoševičius, "Extracting sentence embeddings from pretrained transformer models", *Applied Sciences*, vol. 14, no. 19, p. 8887, 2024.

[19] M. Maslych et al., "Mitigating response delays in free-form conversations with llm-powered intelligent virtual agents", in *Proceedings of the 7th ACM Conference on Conversational User Interfaces*, 2025, pp. 1–15.

[20] F. Zimmermeister, *Whisper large v3 turbo german*, https://huggingface.co/primeline/whisper-large-v3-turbo-german, retrieved: March, 2026, 2024.

[21] A. Radford et al., "Robust speech recognition via large-scale weak supervision", in *International conference on machine learning*, PMLR, 2023, pp. 28 492–28 518.

[22] P. Yu, L. Merrick, G. Nuti, and D. Campos, "Arctic-embed 2.0: Multilingual retrieval without compromise", *arXiv preprint arXiv:2412.04506*, 2024.

[23] F. Mo et al., "History-aware conversational dense retrieval", in *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.

[24] Rhasspy, *Piper: A fast, local neural text to speech system*, https://github.com/OHF-Voice/piper1-gpl, retrieved: March, 2026, 2023.

[25] *Rhasspy/piper-voices · Hugging Face — huggingface.co*, https://huggingface.co/rhasspy/piper-voices, retrieved: March, 2026.

[26] Thorsten Müller, *Tv-44khz-full (revision ff427ec)*, retrieved: March, 2026, 2024. DOI: 10.57967/hf/3290.

[27] C. Muratori, *Immediate-mode graphical user interfaces*, https://caseymuratori.com/blog_0001, retrieved: March, 2026, 2005.

[28] F. S. Marcondes et al., "Using ollama", in *Natural Language Analytics with Generative Large-Language Models: A Practical Approach with Ollama and Open-Source LLMs*, Springer, 2025, pp. 23–35.

[29] H. Brugman, A. Russel, and X. Nijmegen, "Annotating multi-media/multi-modal resources with elan.", in *LREC*, Lisbon, 2004, pp. 2065–2068.

[30] K. Pitsch, "Answering a robot's questions. participation dynamics of adult-child-groups in encounters with a museum guide robot. in: Réseaux, 220-221 (2-3), 113-150, https://doi.org/10.3917/res. 220.0113.", 2020.

[31] W. M. Association et al., "World medical association declaration of helsinki: Ethical principles for medical research involving human subjects", *Jama*, vol. 310, no. 20, pp. 2191–2194, 2013.

[32] *eTIC – electronic Tool for Informed Consent documents*, https://etic.med.tum.de, retrieved: March, 2026, Arbeitskreis Medizinischer Ethik-Kommissionen in der Bundesrepublik Deutschland (AKEK), 2026.

[33] J. Peng et al., "A survey on speech large language models for understanding", *Authorea Preprints*, 2025.

[34] W. Cui et al., "Recent advances in speech language models: A survey", in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025, pp. 13 943–13 970.

[35] C. Wang et al., "Opens2s: Advancing fully open-source end-to-end empathetic large speech language model", in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2025, pp. 906–917.

[36] V. Bhardwaj et al., "Automatic speech recognition (asr) systems for children: A systematic literature review", *Applied Sciences*, vol. 12, no. 9, p. 4419, 2022.

[37] J. Tobin et al., "Automatic speech recognition of conversational speech in individuals with disordered speech", *Journal of Speech, Language, and Hearing Research*, vol. 67, no. 11, pp. 4176–4185, 2024.

[38] J. R. Green et al., "Automatic speech recognition of disordered speech: Personalized models outperforming human listeners on short phrases.", in *Interspeech*, vol. 2021, 2021, pp. 4778–4782.

[39] L. P. Guldimann, "Speech recognition for german-speaking children with congenital disorders: Current limitations and dataset challenges", M.S. thesis, ETH Zurich, 2024.

[40] A. Martin et al., "Project euphonia: Advancing inclusive speech recognition through expanded data collection and evaluation", *Frontiers in Language Sciences*, vol. 4, p. 1 569 448, 2025.

[41] M. Hasegawa-Johnson et al., "Community-supported shared infrastructure in support of speech accessibility", *Journal of Speech, Language, and Hearing Research*, vol. 67, no. 11, pp. 4162–4175, 2024.

[42] L. Klöser, M. Beele, J.-N. Schagen, and B. Kraft, "German text simplification: Finetuning large language models with semi-synthetic data", *arXiv preprint arXiv:2402.10675*, 2024.

[43] A. F. Krause and K. Essig, "Protecting privacy using low-cost data diodes and strong cryptography", in *Intelligent Computing*, K. Arai, Ed., vol. 508, Series Title: Lecture Notes in Networks and Systems, Cham: Springer International Publishing, 2022, pp. 776–788, ISBN: 978-3-031-10466-4 978-3-031-10467-1.

[44] S. Federici et al., "Inside pandora's box: A systematic review of the assessment of the perceived quality of chatbots for people with disabilities or special needs", *Disabil. Rehabil. Assist. Technol.*, vol. 15, no. 7, pp. 832–837, 2020. DOI: https://doi.org/10.1080/17483107.2020.1775313.

[45] J. Abascal and L. Azevedo, "Fundamentals of inclusive hci design", in *International Conference on Universal Access in Human-Computer Interaction*, Springer, 2007, pp. 3–9.

[46] D. Benyon, *Designing user experience*. Pearson UK, 2019.

[47] B. Woods et al., "Cognitive stimulation to improve cognitive functioning in people with dementia", *Cochrane database of systematic reviews*, no. 1, 2023.

Effectiveness of Attribute-Matching Agents on User Impressions and Recommendation Satisfaction in Human-Agent Interactions

Yoshimasa Ohmoto
Faculty of Informatics
Shizuoka University
Shizuoka, Japan
e-mail: ohmoto-y@inf.shizuoka.ac.jp

Reika Goda
Faculty of Informatics
Shizuoka University
Shizuoka, Japan
e-mail: goda.reika.17@shizuoka.ac.jp

***Abstract*—In the realm of human-agent interaction, personalized attribute-matching has emerged as a pivotal factor in enhancing user experiences and satisfaction. This study investigates the effectiveness of attribute-matching agents on user impressions and recommendation satisfaction through a controlled experimental approach using a museum appreciation simulation game. We assessed participants who interacted with agents designed to align attributes with user characteristics, specifically focusing on curiosity and individuality traits. Agents employing attribute-matching received significantly more favorable evaluations across key dimensions: agreement, desire for future interaction, perceived understanding, and appreciation. Notably, the impact was particularly pronounced among participants exhibiting high inquisitiveness and high uniqueness scores. These results highlight the significance of implementing personalized attribute-matching strategies. Conversely, attribute-matching showed limited influence on actual behavioral changes, underscoring the necessity for more dynamic task design.**

Keywords-Human-agent interaction; attribute-matching; personality; recommendation system; personalization.

I. INTRODUCTION

In recent years, artificial intelligence technologies have accelerated research into intelligent agents capable of engaging meaningfully with humans [1]. These agents are conceptualized not merely as task execution tools, but as entities capable of forming significant social bonds with humans [2]. Despite these advancements, users frequently perceive agents as mechanistic constructs rather than entities with distinct personalities [3][4], presenting a considerable obstacle to cultivating enduring human-agent interactions.

This study seeks to integrate principles of human-human interaction into intelligent agents' design. This methodology is grounded in research demonstrating the applicability of social norms governing Human-Human Interaction (HHI) to Human-Computer Interaction (HCI) [5].

Attribute-matching draws from psychological principles of similarity-attraction and homophily, suggesting individuals form stronger connections with others sharing similar characteristics [6][7]. This principle encompasses value similarity and status similarity [8], manifesting in digital contexts and social media platforms [9][10]. Individuals sharing similar experiences or knowledge are more inclined to cultivate interpersonal relationships [11]. When applied to human-agent interaction, this principle suggests users may develop more positive impressions when interacting with agents exhibiting similar attributes.

Two theoretical frameworks guide this research. First, common ground theory [12] suggests successful communication requires shared knowledge and mutual understanding [13]. In computational agents, common ground mechanisms include embodiment, social features, joint action, knowledge base, and mental models, thereby enhancing communication quality and user satisfaction. Second, the Media Equation [14] explains why people treat agents as social actors, responding socially and physiologically to agent cues with emotional-response patterns similar to human interactions [15]. This suggests users unconsciously apply social rules to technological agents, treating them as social entities.

In interpersonal interactions, constructing mental representations of others is imperative for predicting actions and enabling seamless exchanges [16]. Greater similarity between partners facilitates this model formation, leading to smooth and sustainable interactions. Integration of similarity principles into human-agent interaction demonstrates efficacy [3].

We hypothesize that designing agents whose characteristics correspond with users' attributes will significantly improve user impressions and trust. Characteristics include both demographic (e.g., gender, age) and psychological factors (e.g., values, interests). Attribute congruence is intrinsically associated with common ground formation, facilitating enhanced mutual understanding between agents and users. We developed an agent with varied levels of attribute matching and executed interaction experiments entailing action recommendations directed at users. The experiment evaluated users' perceptions of agents and their recommendations, while also analyzing the influence of attribute matching. Furthermore, we explored the correlation between users' personal characteristics and the attribute matching process.

This paper is organized as follows: In Section 2, we briefly introduce related works. Section 3 provides an outline of the proposed method. Section 4 contains a description of our experiment that compared experimental and control groups, and presents our results. In Section 5, we discuss the achievements of this research and some future work. We present our conclusion and future work in Section 6.

 ISBN: 978-1-68558-375-0

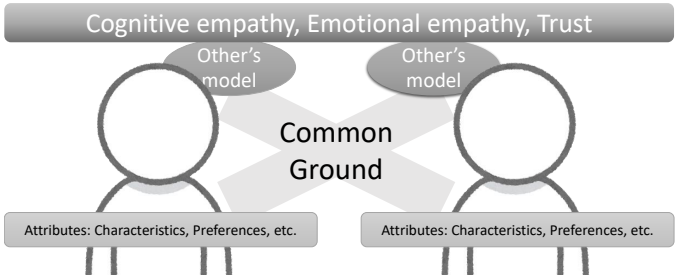


Figure 1. Formation of other's model and positive impression by attribute matching.

II. Related Work

In Human-Agent Interaction (HAI), agent functions have transformed from information providers to interactive communicative partners [17][18]. This transformation is manifest in non-task-oriented dialogue agents facilitating organic social engagement [19]. Recent advancements underscore the significance of individual cognitive models in shaping interactions with artificial agents [20]. Methodologies, such as minimal design [21] and Common Ground theory-informed strategies [22], facilitate enhanced social connections during human-agent exchanges.

Substantial research has examined incorporating human attributes into artificial agents, demonstrating that integrating human characteristics augments interaction consistency and favorably affects user perceptions [23][24]. Agent personality traits exert beneficial influence on user assessments [25]. Equipping agents with distinctive traits enables users to construct 'other person models,' promoting extended and meaningful interactions.

Attribute matching involves static alignment of visible agent traits to create surface-level similarity and establish initial rapport [26][27]. Adaptive personalization involves dynamic modification of agent behavior over time using data-driven user models to continuously refine interaction strategies. While both approaches aim to improve user experience, they operate through different mechanisms. Attribute matching is particularly effective for establishing initial rapport and positive first impressions, while adaptive personalization suits long-term relationship building.

III. Attribute Matching

In this study, we examine attribute matching's impact on user impressions and satisfaction with agent recommendations. We view trust building through forming 'other's models' via attribute matching, which subsequently influences empathy formation (Figure 1). Attitude similarity positively affects interpersonal attraction among humans [28]. When humans perceive personal attributes as common ground through agent similarities, this enhances 'other's model' estimation. Successfully estimating these models fosters affinity and liking toward agents, eventually evolving into empathy and trust.

Importantly, when agents deviate from user expectations, even thoroughly established rapport may be compromised. Thus, agents must perpetually assimilate user characteristics through engagement and respond appropriately. By demonstrating understanding of diverse user characteristics, agents can preserve relationships even amidst conflicts. Virtual agents engineered to cultivate and sustain enduring social-emotional bonds are esteemed and trusted more significantly [29].

In the present study, the agent estimates user characteristics and facilitates model building through self-disclosure and attribute-aligned recommendations. We assist users in recognizing common attributes by offering actions and perspectives congruent with their values.

a) Estimation of User Attributes:: Before interaction, the agent estimates user characteristics using pre-obtained survey data, creating an informed profile that enables personalized communication.

b) Dynamic Attribute Refinement:: During interaction, the agent collects and analyzes user preferences and behaviors. When predictions prove inaccurate, attribute information is immediately adjusted, enabling the agent to refine its understanding iteratively.

c) Self-Disclosure Based on Attribute Alignment:: The agent presents insights grounded in attribute data, adapting to curiosity levels. For highly curious users, it emphasizes unique perspectives; for others, it highlights consensus-based views. This nuanced approach ensures personalized engagement.

d) Action Recommendation Based on Attributes:: When users make choices, the agent recommends actions aligned with both prior survey data and interaction-acquired information, providing relevant and personalized recommendations.

Survey data provides an efficient foundation for prior knowledge. Through interaction, participants disclose values and intentions, which the agent uses to propose aligned actions. By mirroring participant viewpoints, agents facilitate recognition of common attributes. This self-disclosure process cultivates consensus and empathy [30]. As the agent's perspectives increasingly resonate with users' inclinations, user disposition improves, enhancing acceptance of recommendations. This iterative process sustains and intensifies social rapport.

IV. Experiment

This experiment investigated the influence of attribute congruence on users' perceptions of an artificial agent and their overall satisfaction with the recommendations provided by the agent in its capacity as an intermediary. The research incorporated two distinct categories of agents: one configured to utilize attribute matching techniques and the other devoid of such characteristics. The agent employing attribute matching formulated recommendations grounded in initial questionnaires that evaluated the individual traits of the participant, which included the application of the Uniqueness Scale posited by Yamaoka [31] and the inquisitiveness dimension derived from the Critical Thinking Orientation Scale developed by Hirooka et al. [32], along with attributes ascertained during the interactive task. The efficacy of attribute matching was assessed through subsequent questionnaires.

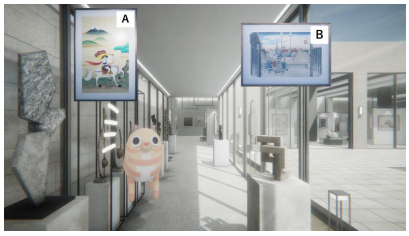


Figure 2. Screen shot of the museum appreciation simulation game.

Figure 3. The appearance of the agents.

A. Task

We employed a "museum appreciation simulation game" developed using Unity platform. Participants explored a virtual museum alongside an agent, examining three artwork categories. Prior to each category, the agent proposed two viewing options based on constrained examination duration. Participants selected one agent-recommended artwork, and the agent elucidated each choice. Figure 2 shows the screen shot of the museum appreciation simulation game.

The attribute-matching agent customized recommendations based on the Uniqueness Scale and inquisitiveness level. For high inquisitiveness participants, it emphasized subjective assessments; for low inquisitiveness, objective evaluations. For high uniqueness participants, it accentuated individuality; for low uniqueness, unity. The control agent provided systematically contrasting recommendations. Following the evaluation of each artwork, participants assessed their satisfaction with the recommended piece utilizing a scale ranging from 1 to 5. Figure 3 shows the appearance of the agents.

B. Experimental Settings

Experiments were conducted in a controlled environment utilizing a 70-inch display (SHARP PN-H701) in conjunction with a Nintendo Joy-Con game controller. Participants were positioned in front of the display, employing the controller to interact with the game. The precise configuration of the experimental setup is illustrated in Figure 4. Engagements with the agent were conducted through the Wizard Of Oz (WOZ) methodology, wherein the experimenter manipulated the agent in accordance with established protocols.

C. Participants

The study involved 19 undergraduate students (15 males, 4 females; M = 22.3, SD = 2.16). Due to technical malfunction, 18 participants' data were analyzed. Participants were classified as high/low based on Uniqueness Scale and inquisitiveness median values.

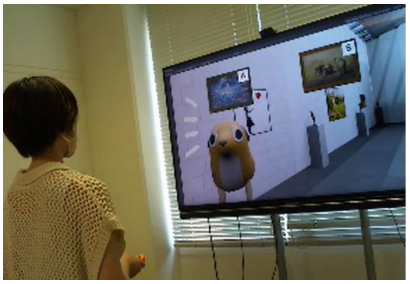


Figure 4. The experimental setting.

D. Procedure

Participants received experimental orientation and trial sessions. Each participant engaged with both attribute-matching and non-attribute-matching conditions, with sequence counterbalanced. Upon completion, participants completed follow-up questionnaires assessing both agents using a 7-point Likert scale for six items: agreement, desire for future interaction, understanding, appreciation, and artwork satisfaction.

Individual assessments utilized the GodSpeed questionnaire, evaluating Anthropomorphism, Animacy, Likability, Perceived intelligence, and Perceived safety [33]. This investigation concentrated on Anthropomorphism, Animacy, and Likability.

E. Results

For direct comparisons among agents, a one-sample Wilcoxon signed-rank test was implemented, whereas ANalysis Of VAriance (ANOVA) was utilized for the assessment of responses garnered from the GodSpeed questionnaire and the satisfaction pertaining to the artworks presented by the agents. Both the attribute-matching and non-attribute-matching conditions functioned as within-participant, with inquisitiveness and uniqueness scales acting as between-participant.

1) Agent Comparative Analysis: Figure 5 illustrates the results obtained from the agent comparison questionnaire, demonstrating that evaluations of the attribute-matching agent were predominantly more favorable across all assessed dimensions. Enhanced assessments for the attribute-matching agent are indicated by a positive value (+1 to +3), while increased ratings for the conventional agent are signified by a negative value (-1 to -3). To evaluate the statistical significance of this discerned trend, a one-sample Wilcoxon signed-rank test revealed significantly higher ratings for the attribute-matching agent across four particular items: "We maintain agreement with one another," "I wish to interact again," "I perceive understanding," and "I feel appreciated" ($p < .05$). These results underscore the notion that a strong alignment between an agents actions and the inherent attributes of the user fosters a deeper sense of familiarity and empathy. Consequently, this alignment not only enhances the overall user experience but also cultivates a more positive perception of the communication process. This suggests that when users feel understood and valued, their willingness to engage with the agent again is significantly increased, reinforcing the importance of attribute matching in the design of interactive agents.

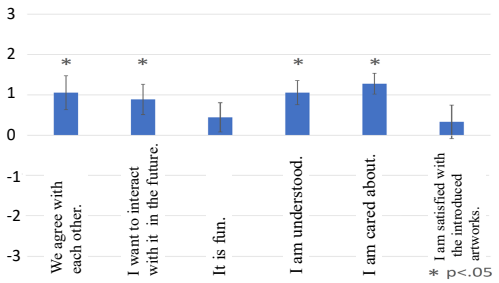


Figure 5. The results from the agent comparison questionnaire.

2) ANOVA: A two-way ANOVA was performed on the outcomes derived from the GodSpeed questionnaire and the degree of recommendation satisfaction. This statistical analysis incorporated high and low categorizations of inquisitiveness and uniqueness scales as between-participant variables, along with attribute matching as a within-participant variable. The findings are illustrated in Figure 6 and Figure 7.

a) Inquisitiveness:

Anthropomorphism: The analysis revealed a marginally significant difference in interaction between inquisitiveness and attribute-matching ($F(1,16) = 3.79$, $p < .1$). To further explore this interaction, a simple main effect test was conducted, which indicated a significant difference in the anthropomorphism score between the groups. Specifically, participants classified in the high inquisitiveness group demonstrated a markedly higher value within the attribute-matching condition ($F(1,16) = 7.15$, $p < .05$). This finding underscores the potential impact of inquisitiveness on the effectiveness of attribute-matching strategies, highlighting the importance of individual differences in cognitive processing.

Animacy: The analysis revealed that there was no statistically significant interaction between inquisitiveness and attribute-matching ($F(1,16) = 2.08$, $p = 0.17$). However, the main effect analysis indicated a trend worth noting: animacy ratings were observed to be marginally significantly higher in the attribute-matching condition ($F(1,16) = 4.41$, $p < 0.1$). This finding suggests that while the interaction between inquisitiveness and attribute-matching was not significant, the attribute-matching condition may still have a noteworthy impact on perceptions of animacy, hinting at the potential importance of this condition in understanding how attributes are matched in relation to perceived animacy.

Likability: There was a significant difference in the interaction between the factors of inquisitiveness and attribute-matching ($F(1,16) = 7.48$, $p < .05$). Further examination through a simple main effect test highlighted that within the attribute-matching condition, participants categorized as having high inquisitiveness demonstrated the higher likability score ($F(1,16) = 4.77$, $p < .05$). This finding underscores the influence of inquisitiveness on outcomes when individuals engage in tasks that require matching attributes. Moreover, within the subset of participants identified as high in inquisitiveness, the results indicated that those in the attribute-matching condition exhibited an even more pronounced effect, yielding a significantly higher score ($F(1,16) = 8.72$, $p < .01$). This further emphasizes the critical role that inquisitiveness plays in enhancing likability in scenarios where attribute-matching is involved.

Artwork Satisfaction: The interaction effect between the variables of inquisitiveness and attribute-matching did not yield a statistically significant result ($F(1,16) = 2.53$, $p = 0.13$). However, when examining the main effects separately, participants categorized in the high inquisitiveness group reported a significantly elevated satisfaction score ($F(1,16) = 7.07$, $p < .05$). This finding underscores the importance of inquisitiveness as a contributing factor to artwork satisfaction, indicating that individuals with higher levels of curiosity tend to derive greater enjoyment or appreciation from the art. Additionally, the attribute-matching condition, while not reaching conventional levels of statistical significance, did show a marginally significant trend with a higher satisfaction score ($F(1,16) = 3.87$, $p < .1$). This suggests that there may be a potential relationship between how well the attributes of the artwork match the viewer's preferences and their level of satisfaction, warranting further investigation to fully understand the nuances of this effect. Overall, these results highlight the complex interplay of inquisitiveness and attribute-matching in shaping individuals' experiences and satisfaction with artwork.

b) Uniqueness Scale:

Anthropomorphism: The interaction between the uniqueness scale and attribute-matching showed a marginally significant difference ($F(1,16) = 3.79$, $p < .1$). To delve deeper, the simple main effects tests indicated that within the attribute-matching condition, participants who were categorized in the high uniqueness scale group exhibited a significantly elevated anthropomorphism score ($F(1,16) = 17.7$, $p < .01$). This finding underscores the importance of uniqueness in enhancing the perception of anthropomorphic qualities. Furthermore, when examining the high uniqueness scale group more closely, it was found that their anthropomorphism scores were also significantly higher when assessed under the attribute-matching condition ($F(1,16) = 7.15$, $p < .05$). Overall, these results highlight the complex interplay between uniqueness and attribute matching in shaping anthropomorphic perceptions.

Animacy: There was no significant difference in the interaction between the uniqueness scale and attribute-matching ($F(1,16) = 1.00$, $p = 0.33$). However, when examining the main effects, noteworthy findings emerged. Specifically, participants in the high uniqueness scale group exhibited a significantly elevated animacy score, as evidenced by the analysis ($F(1,16) = 6.98$, $p < .05$). This result indicates that higher perceived uniqueness is associated with a greater sense of animacy. Furthermore, the animacy score was also significantly higher in the attribute-matching condition ($F(1,16) = 4.15$, $p < .1$), suggesting that when attributes aligned effectively, participants perceived a stronger sense of animacy. These findings highlight the importance of both uniqueness and attribute matching in shaping perceptions of animacy.

Likability: There was no significant difference in the inter-

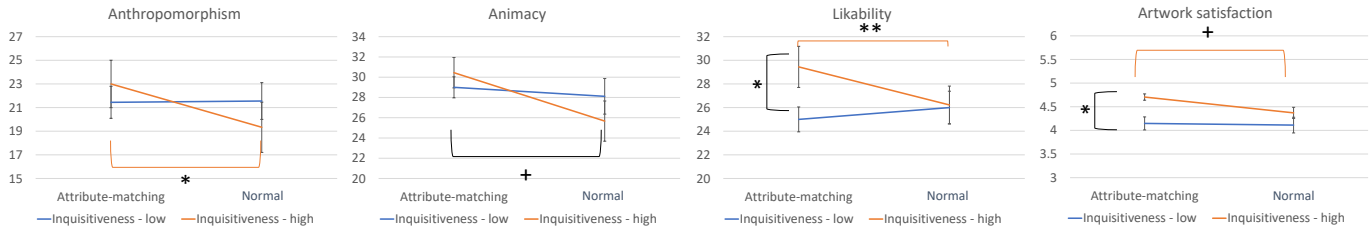


Figure 6. The results of the GodSpeed questionnaire and recommendation satisfaction related to inquisitiveness.

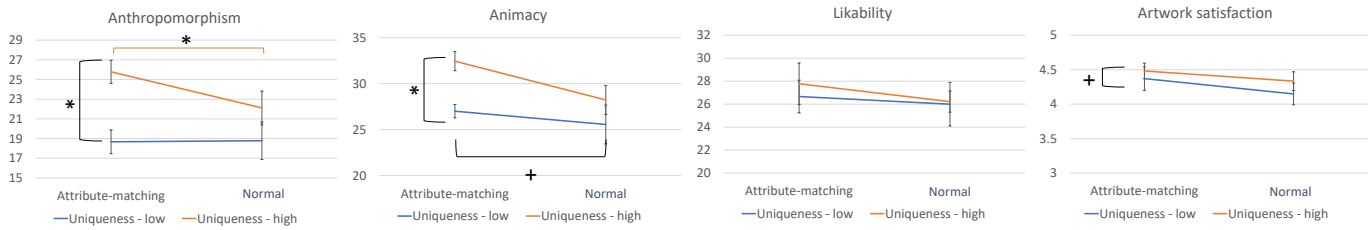


Figure 7. The results of the GodSpeed questionnaire and recommendation satisfaction related to uniqueness.

action between the uniqueness scale and attribute-matching ($F(1,16) = 0.23$, $p = 0.64$). This suggests that variations in the uniqueness scale did not meaningfully influence the relationship with attribute-matching in this particular study.

Artwork Satisfaction: In the artwork satisfaction score, there was no significant difference in the interaction between the uniqueness scale and attribute-matching ($F(1,16) = 0.14$, $p = 0.72$). In the main effects, the value was marginally significantly higher in the attribute-matching condition ($F(1,16) = 3.37$, $p < .1$). This marginal significance suggests that while the difference is not strong enough to be considered statistically conclusive, there is an observable tendency for individuals to report greater satisfaction when the attributes of the artwork align with their expectations or preferences.

These findings suggest that the alignment of attributes significantly affects the participants' perceptions regarding both the agent and the recommendations provided. Specifically, individuals exhibiting elevated levels of inquisitiveness demonstrated a markedly more favorable impression of the agent whose conduct corresponded with their individual attributes. Furthermore, agents that matched attributes augmented participants' perceptions of anthropomorphism and animacy, particularly among those scoring high on the uniqueness scale.

V. Discussion

We investigated attribute-matching expression impacts on human-agent interactions. Findings indicate that agents employing attribute-matching strategies significantly enhance overall participant impressions and cultivate deeper social empathy. This outcome aligns with Media Equation Theory, which posits that individuals respond to computers as social actors [5]. Attribute matching establishes common ground, promoting similarity recognition crucial for relatable and engaging interaction [34].

Our investigation yielded significant insights into personality traits' impacts on user-agent interactions. Participants displaying high inquisitiveness evaluated attribute-matching agents more positively, enhancing their overall perception and increasing recommendation acceptance. Individuals scoring high on uniqueness reported heightened anthropomorphism and animacy when interacting with attribute-matching agents, forming more favorable impressions. These findings align with frameworks suggesting individuals develop stronger connections toward agents exhibiting similar characteristics [35].

Our study acknowledges limitations. Although relatively high satisfaction with recommendations was observed, analysis did not demonstrate statistically significant effects on actual behavioral changes. This lack of behavioral impact likely stems from the inherently passive task nature. In persuasive technology contexts [36], more engaging and active task designs could yield more insightful results. Future research should investigate how attribute-matching dynamics evolve with extended interactions, considering changing user preferences and expectations [29]. The uniform attribute-matching application underscores critical need for developing tailored strategies considering unique user characteristics and preferences.

VI. Conclusion and Future Work

We conducted an empirical examination of agent attribute alignment impacts on human-agent interactions. Results indicate that attribute alignment functions as a fundamental foundation for favorable impression formation and agent rapport establishment. Individual traits, such as curiosity and distinctiveness, significantly influence users' agent perceptions and recommendation satisfaction. These findings highlight promising opportunities for formulating tailored attribute alignment methodologies for improving agent-user rapport development. Ongoing progress in this domain is anticipated to promote more organic and substantive human-agent relationships, ultimately fostering a societal context where humans and agents coexist harmoniously.

REFERENCES

[1] E. Adamopoulou and L. Moussiades, "Chatbots: History, technology, and applications," *Machine Learning with applications*, vol. 2, p. 100 006, 2020.

[2] K. Dautenhahn, "Socially intelligent robots: Dimensions of human–robot interaction," *Philosophical transactions of the royal society B: Biological sciences*, vol. 362, no. 1480, pp. 679–704, 2007.

[3] M. M. De Graaf and B. F. Malle, "People's explanations of robot behavior subtly reveal mental state inferences," in *2019 14th ACM/IEEE international conference on human-robot interaction (HRI)*, IEEE, 2019, pp. 239–248.

[4] M. M. de Graaf, S. Ben Allouch, and J. A. Van Dijk, "Why would i use this in my home? a model of domestic social robot acceptance," *Human–Computer Interaction*, vol. 34, no. 2, pp. 115–173, 2019.

[5] C. Nass and Y. Moon, "Machines and mindlessness: Social responses to computers," *Journal of social issues*, vol. 56, no. 1, pp. 81–103, 2000.

[6] J. J. Van Bavel, L. M. Hackel, and Y. J. Xiao, "The group mind: The pervasive influence of social identity on cognition," *New frontiers in social neuroscience*, pp. 41–56, 2014.

[7] A. J. Bahns, C. S. Crandall, O. Gillath, and K. J. Preacher, "Similarity in relationships as niche construction: Choice, stability, and influence within dyads in a free choice environment.," *Journal of personality and social psychology*, vol. 112, no. 2, p. 329, 2017.

[8] P. F. Lazarsfeld and R. K. Merton, "Friendship as a social process: A substantive and methodological analysis," in *Freedom and control in modern society*, M. Berger, T. Abel, and C. H. Page, Eds., New York: Van Nostrand, 1954, pp. 18–66.

[9] L. M. Aiello et al., "Friendship prediction and homophily in social media," *ACM Transactions on the Web (TWEB)*, vol. 6, no. 2, pp. 1–33, 2012.

[10] B. Mønsted, P. Sapieżyński, E. Ferrara, and S. Lehmann, "Evidence of complex contagion of information in social media: An experiment using twitter bots," *PloS one*, vol. 12, no. 9, e0184148, 2017.

[11] J. Preece and D. Maloney-Krichmar, "Online communities: Design, theory, and practice," *Journal of computer-mediated communication*, vol. 10, no. 4, JCMC10410, 2005.

[12] H. H. Clark and S. E. Brennan, "Grounding in communication," in *Perspectives on socially shared cognition*, L. B. Resnick, J. M. Levine, and S. D. Teasley, Eds., Washington, DC, US: American Psychological Association, 1991, pp. 127–149. DOI: 10.1037/10096-006

[13] I. Kecskes and F. Zhang, "Activating, seeking, and creating common ground: A socio-cognitive approach," *Pragmatics & Cognition*, vol. 17, no. 2, pp. 331–355, 2009.

[14] B. Reeves and C. Nass, "The media equation: How people treat computers, television, and new media like real people and places," *Cambridge, UK*, vol. 10, no. 10, pp. 19–36, 1996.

[15] A. Reuten, M. Van Dam, and M. Naber, "Pupillary responses to robotic and human emotions: The uncanny valley and media equation confirmed," *Frontiers in psychology*, vol. 9, p. 774, 2018.

[16] D. I. Tamir and M. A. Thornton, "Modeling the predictive social mind," *Trends in cognitive sciences*, vol. 22, no. 3, pp. 201–212, 2018.

[17] A. Følstad and P. B. Brandtzæg, "Chatbots and the new world of hci," *interactions*, vol. 24, no. 4, pp. 38–42, 2017.

[18] J. Gao, M. Galley, and L. Li, "Neural approaches to conversational ai," in *The 41st international ACM SIGIR conference on research & development in information retrieval*, 2018, pp. 1371–1374.

[19] S. Roller et al., "Recipes for building an open-domain chatbot," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2021, pp. 300–325.

[20] L. Zhou, J. Gao, D. Li, and H. Y. Shum, "The design and implementation of xiaoice, an empathetic social chatbot," *Computational Linguistics*, vol. 46, no. 1, pp. 53–93, 2020.

[21] N. Matsumoto, H. Fujii, M. Goan, and M. Okada, "Minimal design strategy for embodied communication agents," in *ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, 2005.*, IEEE, 2005, pp. 335–340.

[22] S. Kiesler, "Fostering common ground in human-robot interaction," in *ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, 2005.*, 2005, pp. 729–734. DOI: 10.1109/ROMAN.2005.1513866

[23] K. Isbister and C. Nass, "Consistency of personality in interactive characters: Verbal cues, non-verbal cues, and user characteristics," *International Journal of Human-Computer Studies*, vol. 53, no. 2, pp. 251–267, 2000, retrieved: 2026.03.13, ISSN: 1071-5819. DOI: 10.1006/ijhc.2000.0368

[24] Y. Kim and S. S. Sundar, "Anthropomorphism of computers: Is it mindful or mindless?" *Computers in Human Behavior*, vol. 28, no. 1, pp. 241–250, 2012, retrieved: 2026.03.13, ISSN: 0747-5632. DOI: 10.1016/j.chb.2011.09.006

[25] L. Robert, "Personality in the human robot interaction literature: A review and brief critique," in *Proceedings of the 24th Americas Conference on Information Systems*, 2018, pp. 1–10.

[26] Z. Yu, J. Lian, A. Mahmoody, G. Liu, and X. Xie, "Adaptive user modeling with long and short-term preferences for personalized recommendation.," in *IJCAI*, vol. 7, 2019, pp. 4213–4219.

[27] C. Wu, F. Wu, Y. Huang, and X. Xie, "Personalized news recommendation: Methods and challenges," *ACM Transactions on Information Systems*, vol. 41, no. 1, pp. 1–50, 2023.

[28] D. Byrne, "Interpersonal attraction and attitude similarity.," *The journal of abnormal and social psychology*, vol. 62, no. 3, p. 713, 1961.

[29] T. W. Bickmore and R. W. Picard, "Establishing and maintaining long-term human-computer relationships," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 12, no. 2, pp. 293–327, 2005.

[30] J. A. Bargh, M. Chen, and L. Burrows, "Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action.," *Journal of personality and social psychology*, vol. 71, no. 2, p. 230, 1996.

[31] S. Yamaoka, "The development and validation of the uniqueness scale," *Jpn. J. Soc. Psychol*, vol. 9, pp. 181–194, 1993.

[32] S. Hirooka, "An exploratory study of measurement of "the orientation toward critical thinking"(2)," *Bulletin of Integrated Center for Educational Research and Practice (Mie University)*, vol. 21, p. 93, 2001.

[33] C. Bartneck, D. Kulic, and E. Croft, "Measuring the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots," Jun. 2017. DOI: 10.6084/m9.figshare.5154805.v1

[34] M. J. Pickering and S. Garrod, "Toward a mechanistic psychology of dialogue," *Behavioral and brain sciences*, vol. 27, no. 2, pp. 169–190, 2004.

[35] H. Tajfel, "An integrative theory of intergroup conflict," *The social psychology of intergroup relations/Brooks/Cole*, 1979.

[36] B. J. Fogg, "Persuasive technology: Using computers to change what we think and do," *Ubiquity*, vol. 2002, no. December, p. 2, 2002.