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

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.

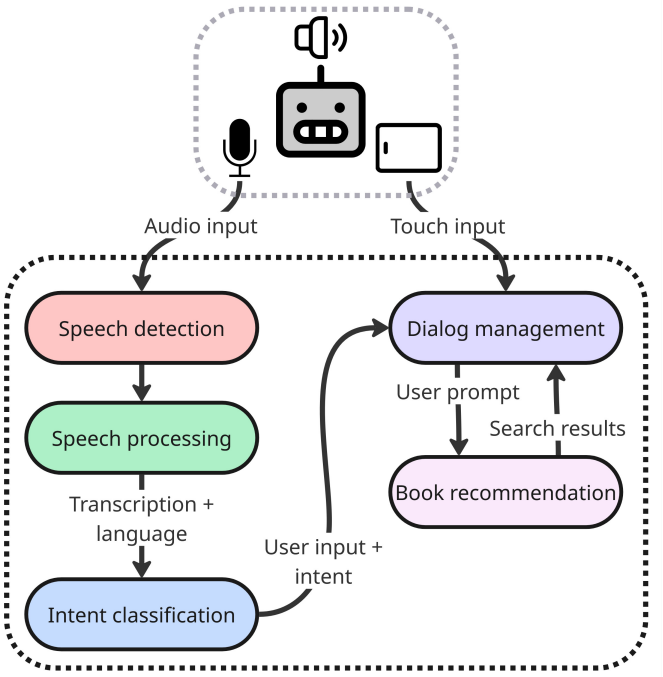

Figure 1. General interaction flow.

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.

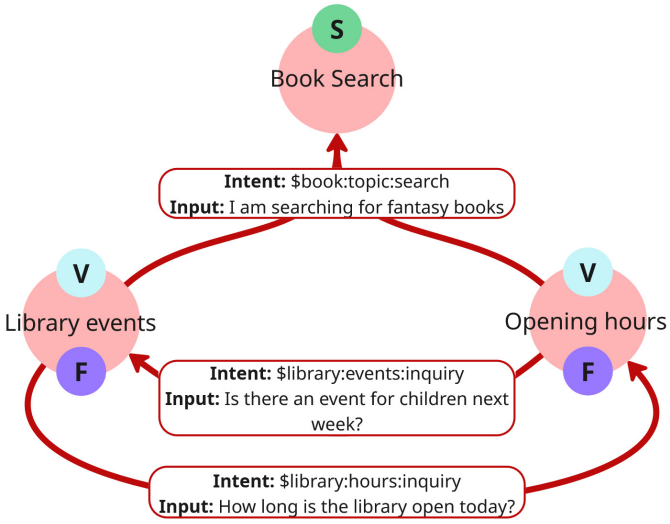

Figure 2. Example dialogue interaction graph with three nodes.

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].