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

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.

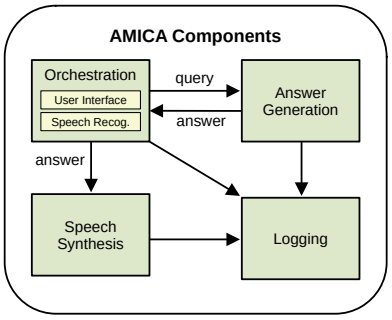

Figure 1. System Architecture. AMICA uses a modular, scalable and distributed architecture consisting of four modules: 1. an orchestration module including user interface and speech recognition sub-modules; 2. an answer generation module; 3. a speech synthesis module and 4. a module for optional logging. 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

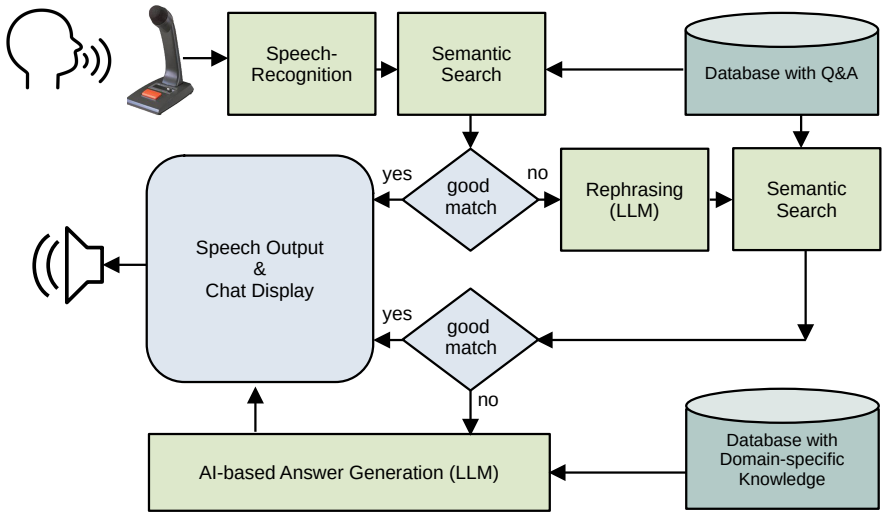

Figure 3. AMICA data-flow diagram. The response to a query is generated in up to three stages. Stage 1 - Retrieval: The question or query of the user is compared with all available questions in the Q&A database using semantic search. If the matching score is high enough, the corresponding answer is directly 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:

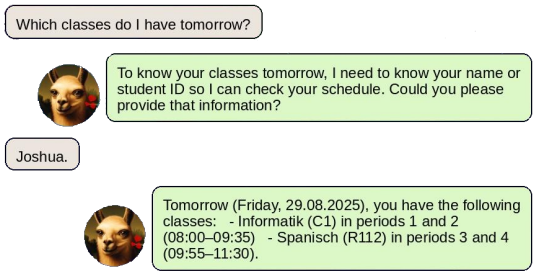

Figure 4. The assistant displays the conversation thread in an established layout similar to the chat history of typical messenger apps.

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

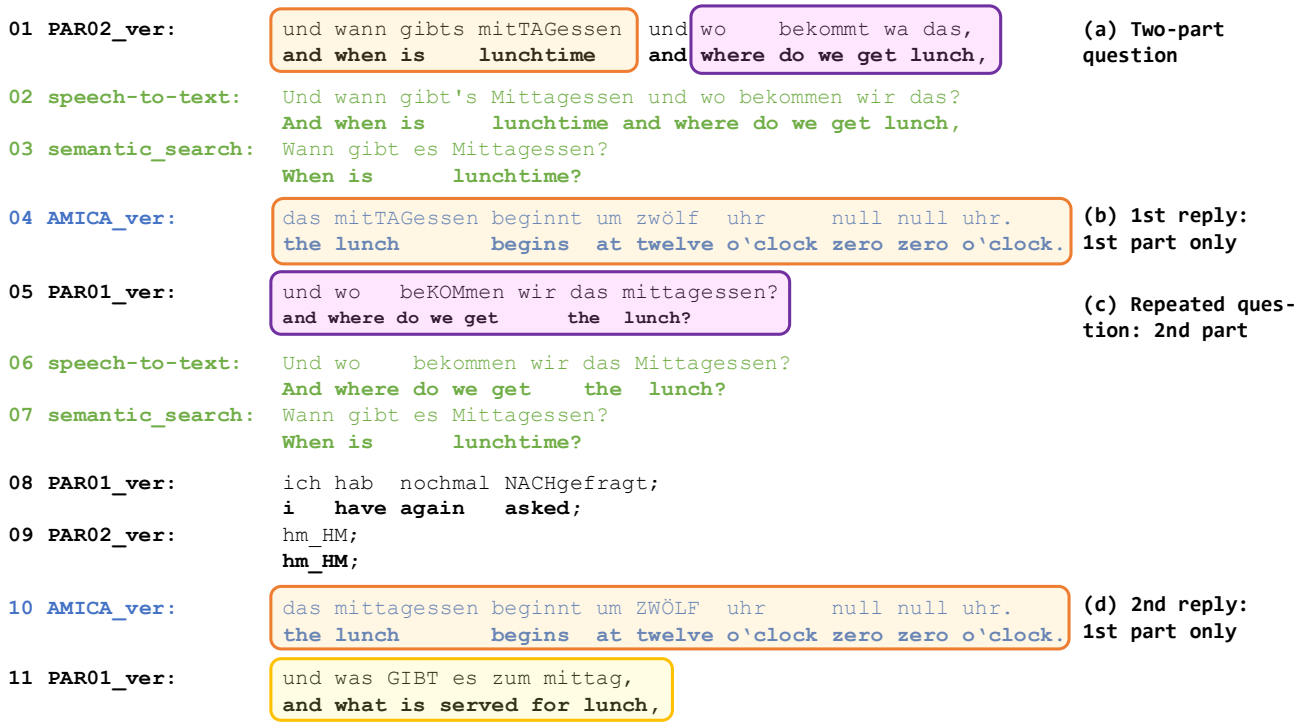

Figure 5. Transcript of verbal actions and log files of one interactional sequence in which two participants (PAR01 and PAR02) use the voice assistant AMICA in a semi-experimental setup.

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.