Implementing an Inclusive Communication System with RAG-enhanced Multilingual and Multimodal Dialogue Capabilities

Cheng-Yun Wu
Institute of Information Management,
National Taipei University
New Taipei City, Taiwan
clever900816@gmail.com

Bor-Jen Chen
Institute of Information Management,
National Taipei University
New Taipei City, Taiwan
s711336103@gm.ntpu.edu.tw

Wen-Hsin Hsiao
Institute of Information Management,
National Taipei University
New Taipei City, Taiwan
s711336107@gm.ntpu.edu.tw

Hsin-Ting Lu
Institute of Information Management,
National Taipei University
New Taipei City, Taiwan
s711436119@gm.ntpu.edu.tw

Yue-Shan Chang
Department of Computer Science and Information Engineering,
National Taipei University
New Taipei City, Taiwan
ysc@mail.ntpu.edu.tw

Chen-Yu Chiang
Department of Communication Engineering,
National Taipei University
New Taipei City, Taiwan
cychiang@mail.ntpu.edu.tw

Chao-Yin Lin
Department of Social Work,
National Taipei University
New Taipei City, Taiwan
cylin@mail.ntpu.edu.tw

Yu-An Lin
Department of Social Work,
National Taipei University
New Taipei City, Taiwan
yuanlin@gm.ntpu.edu.tw

Min-Yuh Day*
Institute of Information Management,
National Taipei University
New Taipei City, Taiwan
myday@gm.ntpu.edu.tw

***Abstract*—Inclusive communication technologies are essential for ensuring equal access to information and public services among individuals with speech impairments and speakers of low-resource languages. While large language models and multimodal systems have made significant progress, current solutions often lack domain-specific support and cross-lingual flexibility. This study aims to address this gap by developing a lightweight RAG-enhanced dialogue system that integrates multimodal input and multilingual generation for inclusive task-oriented scenarios. The system architecture combines vector-based semantic retrieval and generative AI to support text speech and image inputs across mobile and web platforms. It was co-designed with assistive organizations and deployed iteratively in three stages, improving scalability accessibility and response quality. Results from preliminary testing show increasing engagement and positive feedback on system usability content accuracy and interaction style. This research contributes a practical framework for inclusive AI communication support with implications for accessibility-focused system design in both social and public service domains.**

Keywords—Retrieval-Augmented Generation (RAG), Multimodal Dialogue System, Multilingual Communication, Inclusive AI, Non-Profit Organization

I. Introduction

Communication is a fundamental human right. However, millions of people around the world, especially those with speech impairments or who speak low-resource languages, still face difficulties accessing digital information and services. For these individuals, daily tasks such as seeking help or receiving public information can be limited by language barriers and inaccessible user interfaces. To promote equity in education, healthcare, and public participation, there is an urgent need to develop inclusive communication technologies. Task-Oriented Dialogue Systems (TODs), which offer structured support for completing specific tasks, provide a promising solution for meeting diverse communication needs in areas such as customer service, health consultations, and public assistance [1].

Although assistive communication tools have gradually improved, most systems still lack the flexibility needed to adapt to real-world communication scenarios. Many rely on fixed response templates or machine translation, which limits their effectiveness in supporting real-time and cross-cultural dialogue. The emergence of large language models (LLMs) has enhanced the naturalness and contextual understanding of AI-generated conversations. In parallel, multimodal dialogue systems have expanded input options to include speech, images, and text, improving accessibility and user experience.

However, LLMs still tend to generate inaccurate or fabricated content, especially when they lack access to reliable external knowledge [2].To address this issue, recent studies have proposed the use of Retrieval-Augmented Generation (RAG), which combines LLMs with vector-based retrieval from external knowledge sources. This hybrid approach improves factual accuracy and contextual relevance, making it a promising solution for knowledge-intensive dialogue tasks [3].

Despite advancements in LLMs and RAG, few systems have effectively integrated these technologies with multimodal processing and multilingual understanding [4, 5]. This lack of integration limits their applicability in inclusive communication settings, where systems must handle visual, auditory, and linguistic diversity simultaneously. Therefore, it is essential to design dialogue systems that can process various input formats and provide accurate, accessible responses in multiple languages.

This study aims to bridge the above gap by developing a lightweight and scalable RAG-enhanced task-oriented dialogue system that supports multimodal and cross-lingual

communication for users with diverse abilities. The goals of this research are: (1)To build a multimodal RAG-based dialogue system that supports text, image, and speech input, improving accessibility for users with speech impairments or multilingual needs.(2)To enable cross-lingual and cross-device interactions, including real-time support on web platforms, mobile devices, and assistive tools.(3)To collaborate with assistive technology organizations to ensure the system addresses the practical needs of underserved communities.

II. Literature Review

A. Inclusive Communication and Assistive Dialogue Systems

Inclusive communication is essential to ensuring equal access to information and participation for individuals with disabilities or communication barriers. As emphasized in global frameworks such as the UN Convention on the Rights of Persons with Disabilities (CRPD), accessibility to information and communication technologies is a fundamental right, and removing such barriers is a prerequisite for full societal inclusion [6]. Assistive technologies play a vital role in supporting this goal.

Recent studies have explored how artificial intelligence (AI) and interactive technologies can enhance inclusive communication. AI-powered tools such as adaptive learning systems, generative chatbots, and virtual assistants have shown potential in creating personalized, accessible experiences for users with diverse needs [7]. Also, integrating language models like ChatGPT with assistive interfaces has opened new possibilities for flexible, real-time communication support [8].

B. Task-Oriented Dialogue Systems (TOD)

TODs are designed to help users accomplish specific tasks, such as booking services, making inquiries, or providing guidance in structured domains. Unlike open-domain chatbots that focus on conversational fluency, TODs prioritize task completion and accurate information exchange [1].

Traditional TODs typically adopt a modular pipeline architecture with components such as natural language understanding (NLU), state tracking, dialogue policy, and response generation. These systems are interpretable and easier to control, making them suitable for practical deployment. Recent work has also explored end-to-end models that train all components jointly, offering improved adaptability while reducing engineering complexity [9].

Building on these developments, hybrid TOD systems now incorporate LLMs and vector retrieval to enhance flexibility and reduce reliance on annotated data. Such systems leverage LLMs for zero-shot intent detection, entity extraction, and dialogue policy generation, while vector databases improve semantic matching and context awareness [10].

C. Maintaining the Integrity of the Specifications

LLMs, such as GPT, have demonstrated strong capabilities in NLU and natural language generation (NLG), enabling applications like question answering, summarization, and dialogue generation [11]. However, LLMs sometimes produce incorrect or misleading content, known as hallucinations. These issues often result from limited training data, unstable inference, and missing access to up-to-date external knowledge [12].

To address these issues, RAG has emerged as an effective approach. By combining a retriever module with a generator, RAG enables LLMs to access external knowledge sources, improving factuality, transparency, and adaptability [3]. Recent RAG models use different ways to mix retrieved content with the input, such as through prompts or hidden layers [13]. This hybrid architecture reduces hallucination and enhances domain-specific reliability, making it suitable for knowledge-intensive dialogue systems.

D. Multimodal Dialogue Systems

Multimodal dialogue systems are designed to understand and generate responses across multiple input types, such as text, images, and other sensory data. With the rise of large multimodal models (LMMs), recent research has focused on building unified architectures that can handle a variety of tasks within a single framework. These models aim to support general-purpose capabilities like visual question answering, image captioning, and text-image reasoning, thereby enhancing their effectiveness in diverse multimodal applications [14, 15]. Although benchmarks for integrated multimodal capabilities are emerging, such as MM-Vet, the evaluation of LMMs in real-world settings remains an open area for further exploration [16].

E. Cross-lingual and Multilingual Dialogue Models

Building dialogue systems that support multiple languages poses challenges such as limited annotated data and inconsistent performance across typologically diverse languages. Existing studies highlight that multilingual task-oriented dialogue models often rely on machine translation or multilingual pretrained models like mBERT and XLM-R, which may not generalize well to low-resource languages [17].

One effective approach is intermediate fine-tuning on multilingual data before task-specific training, which has been shown to improve dialogue state tracking, especially in low-resource scenarios [18]. This is particularly relevant for systems involving languages like Vietnamese, where task-specific corpora are scarce.

Additionally, parallel dialogue corpora such as XDailyDialog offer consistent multi-language structures that support evaluation and development across languages [19]. While not covering all language pairs, such resources provide useful guidance for cross-lingual system design.

While prior studies have demonstrated the potential of LLMs, RAG, and multimodal or multilingual dialogue systems across various tasks, several gaps remain. Most existing systems are developed for high-resource languages or general-purpose applications, with limited attention to inclusive, domain-specific, and cross-lingual use cases involving underrepresented languages such as Vietnamese. In addition, few works integrate multimodal, task-oriented, and multilingual capabilities within a unified framework designed for real-world communication support. This study aims to address these gaps by developing a lightweight, RAG-enhanced, cross-lingual, and multimodal dialogue system tailored for inclusive communication scenarios.

III. Methodology

This section outlines the technical framework and development process of the proposed RAG-enhanced

multimodal cross-lingual dialogue system. We begin by presenting the overall system architecture and core technologies. Then, we describe our collaborative approach with assistive organizations, followed by the customization strategies adopted for each use case. Finally, we detail the development process across three system versions, demonstrating the evolution from a basic text-based chatbot to an advanced multimodal assistant with voice and image capabilities.

A. System Overview and Development Framework

1) Overall system architecture: The system architecture, shown in Fig. 1, consists of two main layers: the application environment for user interaction and the conversational AI agent for backend processing. User queries—via text or speech—are encoded using a multimodal embedding model and matched against a vector database built from organizational knowledge sources. Relevant content is retrieved through re-ranking and filtering, then combined with the original query to form an augmented input. This input is passed to a generative model to produce a natural language response. The modular design enables flexible support for multimodal and multilingual communication needs, with integration across LINE and web interfaces.

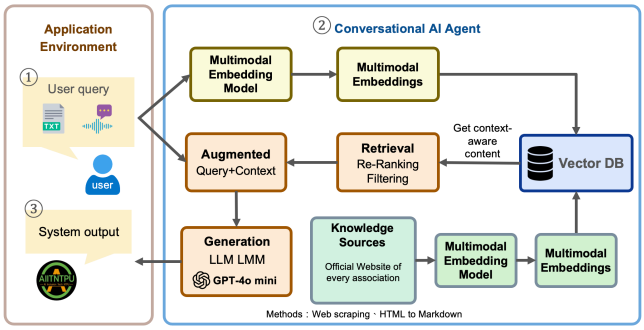


Fig. 1. System architecture of the RAG-enhanced multimodal cross-lingual dialogue system

2) Technology stack: The system was developed through three progressive stages. In the first stage, the prototype was implemented locally using Python and Flask, integrating SentenceTransformer embeddings with FAISS for retrieval, and OpenAI's GPT-4o-mini API for response generation. LINE was used as the primary user interface. In the second stage, the system was deployed to the cloud environment via Google Cloud Run, and speech functionality was added using the Whisper model, maintaining LINE as the interaction platform. In the third stage, we adopted the multilingual-e5-base model for embedding, OpenAI's GPT-4.1 API for generation, and the NVIDIA NIM API with the Gemma-3 27B model for image-to-text processing. The interface was constructed using Gradio to support a flexible, web-based, and multimodal user experience.

B. Organizational Collaboration and Data Collection

1) Partner associations: To ensure real-world applicability and inclusiveness, this study was co-designed in collaboration with five non-profit organizations in Taiwan: the Cerebral Palsy Association of R.O.C., Taiwan Motor Neuron Disease Association (MNDA), the Sunshine Social Welfare Foundation, Good Friend Mission (GFM), and Taiwan Access for All Association. These partners were selected for their direct service to individuals with speech and physical impairments, providing diverse perspectives and communication needs across user demographics. In addition to these organization-specific systems, we also developed an integrated dialogue system that explains the overall research project and supports general inquiries. In total, six systems were implemented to address both individual and collective communication needs.

2) User requirements from partner associations: User requirements was conducted through direct communication between our development team and staff members—including social workers—from each participating organization. During these consultations, team members engaged in semi-structured discussions to identify real-world communication needs, service workflows, and user scenarios specific to each group.

C. Custom Dialogue System Design per Organization

1) Use case adaptation: Each partner organization required tailored interaction designs based on their service content and user preferences. For example, the chairman of MNDA proposed incorporating a disease-specific assessment form into the system. Once completed by the user, the system would organize and store the results in a structured format, facilitating follow-up actions by the association's staff. GFM, which focuses on youth employment and career counseling, emphasized the need for a conversational interface that could provide accurate and context-aware responses to questions raised by both staff and service users. Feedback highlighted that the system should be more relevant and aligned with their internal workflows than general-purpose tools such as ChatGPT. These diverse requirements necessitated the development of organization-specific intents and interaction flows within the system.

2) Knowledge base construction: The knowledge base was constructed by collecting domain-specific content from the websites, internal documents, and interviews provided by partner organizations. The raw text was preprocessed through content cleaning and segmentation to form a structured corpus suitable for downstream processing. To enable semantic retrieval, each segment was encoded into vector representations and indexed using FAISS. During interaction, user queries—whether typed or transcribed from speech—are similarly transformed and compared against the indexed knowledge base to retrieve the most relevant segments. These segments are then combined with the user query and passed to the generative model to produce informative and contextually appropriate responses. This architecture supports efficient, scalable knowledge access and ensures that responses are grounded in authoritative organizational content while remaining adaptive to diverse user inputs.

D. Iterative Development Stages

1) First version: LINE bot prototype for basic text communication: The initial prototype was developed as a LINE chatbot to demonstrate fundamental functionality within a widely used mobile interface. This version supported only text-based queries in Mandarin Chinese. It was implemented locally using Python and Flask, with vector search handled through FAISS and responses generated via OpenAI's GPT-4o-mini API. The goal of this version was to validate core interaction flows and assess user receptiveness to AI-assisted communication. Feedback from early testing helped refine the system's conversational tone, identify frequently asked questions, and confirm the usability of the LINE interface for the target user groups.

2) Second version: bilingual and voice-enabled LINE bot: To address user feedback regarding accessibility, the second version extended the chatbot with automatic speech recognition using Whisper-v1 and introduced bilingual support for Mandarin and English. The system was migrated to a cloud environment via Google Cloud Run, ensuring higher availability and scalability. While continuing to use LINE as the main interface, this version introduced a speech input button to facilitate hands-free interaction, particularly useful for users with motor impairments.

In the early stage of this version, response generation was handled by GPT-4o-mini, and the system did not include any form of interaction logging. In later updates, the generation module was replaced with the more efficient Mistral-small-latest model, and a logging mechanism was introduced. This log captures user inputs, retrieved context segments, and generated responses for backend analysis, error diagnosis, and continuous system refinement. This enhancement provided greater visibility into system behavior, enabling more informed adjustments to both retrieval and generation components.

3) Third version: web-based Gradio multimodal assistant: In the third version, we developed a web-based multimodal assistant using Gradio, enabling richer and more flexible interactions. The system integrated three input modalities—text, speech, and image—expanding support for diverse communication needs and user preferences. Text inputs are embedded and semantically matched with the knowledge base via FAISS, with context passed to GPT-4.1 for response generation. Audio inputs are transcribed using Whisper and processed through the same retrieval-generation pipeline. For image input, the system employs the NVIDIA NIM API with the Gemma-3 27B model to generate descriptive prompts in Traditional Chinese. The Gradio interface clearly separates each modality and delivers immediate, AI-generated responses, ensuring accessible and seamless interactions. This version demonstrates substantial progress in multimodal AI integration, RAG-based knowledge grounding, and cross-lingual support.

IV. Experimental Results

This section presents the evaluation outcomes of the proposed RAG-enhanced multimodal dialogue system developed across three progressive stages. The evaluation includes both qualitative insights, such as representative dialogue outputs, and quantitative results, such as message usage statistics and feedback analysis. To contextualize the system's implementation, we first provide an overview of the partner organizations involved in system co-design, followed by interaction examples and usage statistics.

A. Partner Organizations and Service Domains

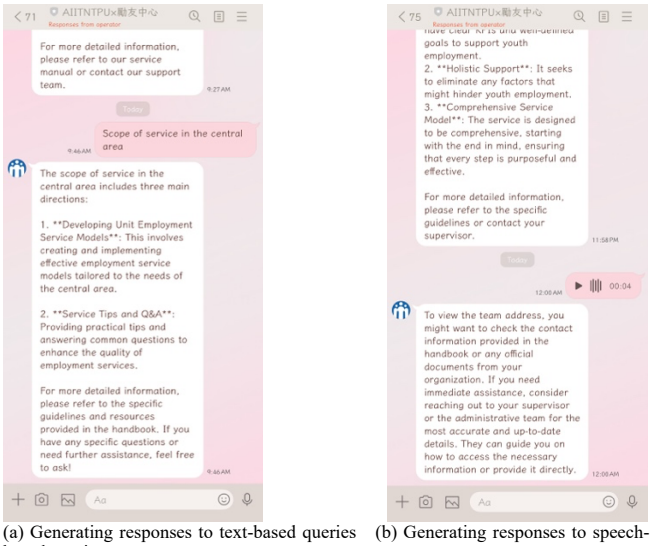
The system was collaboratively developed with five non-profit organizations in Taiwan, each serving distinct user groups with communication or mobility impairments. Table I summarizes the core service focus of each organization and the specific communication needs addressed by the dialogue system during customization and deployment.

TABLE I. Partner Organizations and Their Service Focus

Organization Name	Service Focus
Cerebral Palsy Association of R.O.C.	Support for individuals with cerebral palsy
Taiwan Motor Neuron Disease Association (MNDA)	ALS and motor neuron disease support
Sunshine Social Welfare Foundation	Burn injury rehabilitation and reintegration services
Good Friend Mission (GFM)	Youth employment and social reintegration support
Taiwan Access for All Association	Accessibility advocacy and inclusive policy promotion

B. System Outputs

To demonstrate the system's evolution, we present representative screenshots and interaction examples from the second and third development stages. These include selected dialogue flows from the second-stage LINE bot (Fig. 2), and a multimodal, multilingual dialogue session from the third-stage Gradio system (Fig. 3). Fig. 2 consists of two subfigures: (a) shows an English text query submitted through the LINE bot for GFM, and (b) illustrates a speech-based interaction processed via automatic speech recognition. Subfigure (a) appears on the left and subfigure (b) on the right. The third-stage system (Fig. 3) supports inputs in multiple formats—text, speech, and images—and multiple languages, providing a flexible and accessible user experience. It also functions as an integrated assistant, capable of answering queries related to all subprojects and partner organizations.



(a) Generating responses to text-based queries (b) Generating responses to speech-based queries

Fig. 2. Sample LINE bot interactions with GFM: (a) text query and response; (b) voice input processed via speech recognition

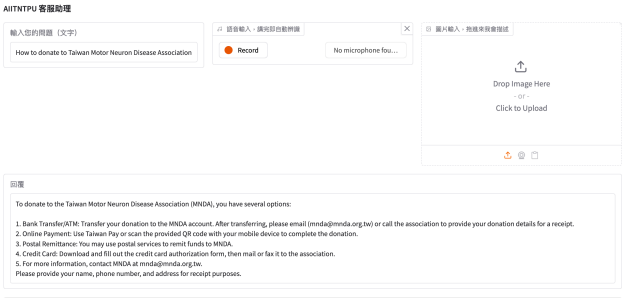


Fig. 3. Multilingual multimodal dialogue in the third-stage Gradio-based system, including query methods via text, speech, and images

C. Usage Statistics

1) LINE bot usage logs: To illustrate system activity, we collected message statistics from the deployed LINE bot over a period of approximately one year. The data, shown in Fig. 4, represent the monthly count of system responses to user

queries. While the overall trend shows fluctuations, we observed that message volume often spiked during periods of active development, testing, and content refinement — suggesting increased internal usage and engagement during system adjustment cycles. Notably, peaks in message responses were followed by gradual stabilization, reflecting the iterative nature of deployment and optimization. In more recent months, message counts have again shown signs of growth, indicating renewed engagement and system readiness for broader adoption.

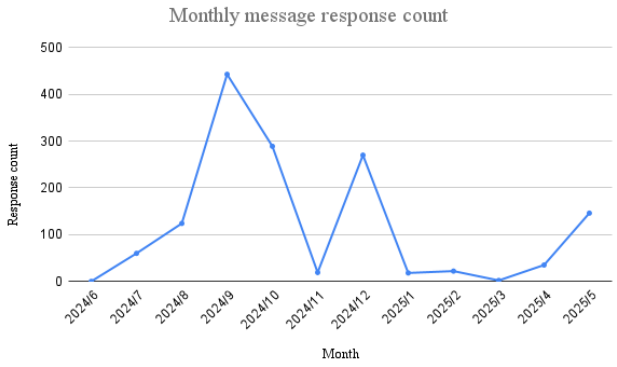


Fig. 4. Monthly message response count from the deployed LINE bot between June 2024 and May 2025

2) Preliminary testing and user feedback: Initial testing conducted by internal teams and partner organization members provided encouraging results, confirming that the system is capable of handling basic inquiries and delivering coherent responses. They offered constructive feedback based on their practical interactions with the system. They recognized the system's potential to serve as a meaningful communication tool and suggested several enhancements to further improve its effectiveness. These included enriching the content base to increase the diversity and depth of responses, as well as incorporating organization-specific information — such as an introduction to the Good Friend Mission — to ensure contextual accuracy. They also expressed interest in refining the tone of the system's replies to make them warmer and more engaging, while maintaining clarity. This feedback affirms the system's foundational utility and provides clear direction for iterative improvement in both knowledge integration and user experience design.

V. Conclusion

This study presents a lightweight, scalable, and RAG-enhanced multimodal dialogue system designed to support inclusive communication for individuals with speech impairments and multilingual needs. By integrating semantic retrieval with large language models, the system advances the application of RAG in the assistive technology domain. It demonstrates how cross-lingual and multimodal capabilities can be effectively embedded into a unified dialogue framework to accommodate diverse input types and linguistic contexts. Furthermore, the three-stage development process illustrates how such systems can be adapted and scaled to serve multiple organizational use cases in real-world environments.

The research offers valuable academic contributions in three areas. First, it demonstrates the practical integration of RAG architecture within assistive, task-oriented dialogue systems. Second, it introduces a modular and extensible design that can be adapted to serve speech-impaired users across various interaction channels. Third, it provides empirical insights into multilingual and multimodal system implementation, which can inform future developments in inclusive dialogue research.

From a managerial perspective, this work highlights the practical value of AI-powered communication systems in enhancing service delivery for non-profit organizations. It enables frontline staff and service users to access information more efficiently, particularly in multilingual or low-resource settings. In addition, the collaboration model established with partner organizations serves as a blueprint for participatory system design in accessibility-focused innovation initiatives.

Future work will focus on expanding system capabilities in three key directions. First, we plan to enhance log recording mechanisms to support deeper evaluation and ongoing system refinement. Second, we aim to extend language support to include additional low-resource and regionally relevant languages, such as Hokkien. Finally, broader deployment across more service contexts and mobile platforms will be pursued to further validate the system's inclusivity and scalability.

Acknowledgment

This work was supported by National Science and Technology Council, Taiwan, under grants NSTC 113-2425-H-305-003-, NSTC 114-2425-H-305-003- and National Taipei University (NTPU), Taiwan and ATEC Group under grants NTPU-112A413E01, and National Taipei University (NTPU), Taiwan under grants 114-NTPU_ORDA-F-004.

References

[1] Z. Zhang, R. Takanobu, Q. Zhu, M. Huang, and X. Zhu, "Recent advances and challenges in task-oriented dialog systems," *Science China Technological Sciences,* vol. 63, no. 10, pp. 2011-2027, 2020.

[2] 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.

[3] P. Lewis *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in Neural Information Processing Systems,* vol. 33, pp. 9459-9474, 2020.

[4] S. Kumar *et al.*, "Bridging the Gap: Dynamic Learning Strategies for Improving Multilingual Performance in LLMs," *arXiv preprint arXiv:2405.18359,* 2024.

[5] N. Chirkova, D. Rau, H. Déjean, T. Formal, S. Clinchant, and V. Nikoulina, "Retrieval-augmented generation in multilingual settings," *arXiv preprint arXiv:2407.01463,* 2024.

[6] A. Altinier, E. Oncins, G. Sauberer, and T. Mehigan, "Demystifying digital accessibility and fostering inclusive mindsets. Compliance with the European standard for digital accessibility EN 301 549," in *European Conference on Software Process Improvement*, 2022: Springer, pp. 595-609.

[7] X. Mitre and M. Zeneli, "Using AI to Improve Accessibility and Inclusivity in Higher Education for Students with Disabilities," in *2024 21st International Conference on Information Technology Based Higher Education and Training (ITHET)*, 2024: IEEE, pp. 1-8.

[8] Z. S. Pomare, G. Hossain, D. S. Maguluri, and G. Prybutok, "ChatGPT as Assistive Technology: Opportunities and Challenges," in *2024 IEEE International Conference on Contemporary Computing and Communications (InC4)*, 2024, vol. 1: IEEE, pp. 1-5.

[9] T.-H. Wen *et al.*, "A network-based end-to-end trainable task-oriented dialogue system," *arXiv preprint arXiv:1604.04562,* 2016.

[10] R. He, J. Gao, J. Guo, and C. Lin, "Research and Implementation of Task-oriented Dialogue Systems Based on Large Language Models and Vector Retrieval," in *2024 4th International Conference on Electronic Information Engineering and Computer Communication (EIECC)*, 2024: IEEE, pp. 1351-1357.

[11] A. F. Mohammad, B. Clark, and R. Hegde, "Large language model (LLM) & GPT, A monolithic study in generative AI," in *2023 Congress in Computer Science, Computer Engineering, & Applied Computing (CSCE)*, 2023: IEEE, pp. 383-388.

[12] L. Huang *et al.*, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *ACM Transactions on Information Systems,* 2023.
[13] S. Wu *et al.*, "Retrieval-augmented generation for natural language processing: A survey," *arXiv preprint arXiv:2407.13193,* 2024.
[14] Q. Sun *et al.*, "Emu: Generative pretraining in multimodality," *arXiv preprint arXiv:2307.05222,* 2023.
[15] B. Zhou *et al.*, "Tinyllava: A framework of small-scale large multimodal models," *arXiv preprint arXiv:2402.14289,* 2024.
[16] W. Yu *et al.*, "Mm-vet: Evaluating large multimodal models for integrated capabilities," *arXiv preprint arXiv:2308.02490,* 2023.
[17] E. Razumovskaia, G. Glavas, O. Majewska, E. M. Ponti, A. Korhonen, and I. Vulic, "Crossing the conversational chasm: A primer on natural language processing for multilingual task-oriented dialogue systems," *Journal of Artificial Intelligence Research,* vol. 74, pp. 1351-1402, 2022.
[18] N. Moghe, M. Steedman, and A. Birch, "Cross-lingual intermediate fine-tuning improves dialogue state tracking," *arXiv preprint arXiv:2109.13620,* 2021.
[19] Z. Liu *et al.*, "XDailyDialog: a multilingual parallel dialogue corpus," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers*, 2023: Association for Computational Linguistics, pp. 12240-12253.