Generative AI in Multimodal Cross-Lingual Dialogue System for Inclusive Communication Support

Vidhya Nataraj
Smart Healthcare Management
National Taipei University
New Taipei City, Taiwan
vidhyanataraj99@gmail.com

Wen-Hsuan Liao
Graduate Institute of Information Management
National Taipei University
New Taipei City, Taiwan
s711136109@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*
Graduate Insitute of Information Management
National Taipei University
New Taipei City, Taiwan
myday@gm.ntpu.edu.tw

Abstract

Advancements in natural language processing have enhanced dialogue systems, making them vital for inclusive technology that facilitates accessible interactions across diverse user needs. However, existing systems often struggle with multimodal inputs, multilingual support, and generating contextually appropriate responses in data-scarce environments. This research addresses these gaps by developing an integrated dialogue system leveraging generative AI models like ChatGPT and multimodal inputs like text, audio, and image. The system utilizes transfer learning and large language models (LLMs) to process multilingual data, generating comprehensive responses tailored to user context. The proposed approach constructs a multimodal cross-lingual task-oriented dialogue system capable of understanding and responding to users in multiple languages and modalities. The proposed multimodal cross-lingual task-oriented dialogue system will enhance functionality and inclusivity compared to traditional unimodal or single-language dialogue systems in providing inclusive communication support. The major research contribution of this study highlights the potential of generative AI in developing accessible dialogue systems that cater to diverse user needs to advance inclusive technology. Practitioner implications of this paper highlight the potential of multimodal cross-lingual dialogue system to foster digital inclusion and inclusive communication support, improving accessibility and equity in human-computer interactions for diverse users.

Keywords: Generative AI, Large Language Models (LLMs), Multimodal, Cross-lingual, Dialogue System, Inclusive Communication Support

I. INTRODUCTION

Care and rehabilitation play crucial roles in enhancing the quality of life for individuals with physical or mental injuries. These efforts also carry significant financial implications, both short-term and long-term, for individuals, their families, and broader societal structures including the healthcare system. The integration of assisted living technologies with intelligent interfaces and advanced dialogue systems in homes and clinics can significantly improve individual quality of life by providing essential support through technology[1]. Building conversational intelligence systems remains a cornerstone objective in the realm of Natural Language Processing (NLP)[2][3]. The capability for machines to interact naturally and seamlessly with humans has not only revolutionized human-machine communication but also led to transformative applications across a wide array of industries. From well-known virtual assistants like Siri and Xiaoice to advanced search and interaction platforms such as the New Bing and Google Bard, dialogue systems have become integral to our digital experience[4]. At the core of this evolution are dialogue systems designed to facilitate interactions that are characterized by attributes such as harmlessness, helpfulness, trustworthiness, and a high degree of personalization [5]–[8]. These systems are engineered to emulate human-to-human conversation, thereby enhancing user experiences, streamlining complex tasks, and providing personalized assistance across diverse domains such as

2835-5776/24/$31.00 ©2024 IEEE
DOI 10.1109/IRI62200.2024.00051

customer support, virtual assistants, healthcare, and education [9]. Task-Oriented Dialogue (TOD) system specifically concentrates on assisting users in achieving explicit tasks or goals, such as making reservations or booking tickets. By adeptly detecting user intentions, tracking dialogue states, making suitable actions, and responding appropriately, TODs serve as highly efficient virtual assistants that deliver relevant information and guidance [10][3].

However, the revolutionary advances in Language Models (LMs), particularly with the advent of LLMs such as ChatGPT, have sparked a transformative evolution in the foundational structure of dialogue systems[11][12]. These advanced models do not merely predict word sequence probabilities; they absorb and interpret extensive world knowledge from their training corpora, enabling them to generate responses that are contextually relevant, deeply nuanced, and highly useful[13][14]. This evolution is particularly pivotal as it catalyzes the integration of advanced capabilities, such as multilingual support and multimodal interactions, thereby broadening the accessibility and applicability of dialogue systems. On the other hand, recent advances in multilingual pre-trained language models (mPLMs) [15]–[17]have conceptually enabled cross-lingual transfer between any two or more languages seen at pretraining or even to unseen languages [18]–[20].

Despite significant progress, there is a notable scarcity of research focused on systems that support cross-language, multimodal TOD. Such systems are crucial for fostering inclusive communication that transcends linguistic and sensory barriers, accommodating diverse user needs through the integration of visual, textual, and auditory data. The development of these comprehensive systems represents a critical area of research. It highlights the urgent need for methodologies that leverage the synergy between LLMs, multimodal interfaces, Generative AI, and multilingual capabilities. These technologies combined have the potential to create dialogue systems that are not only accessible to a broader range of users but are also capable of more deeply personalized and contextually aware interactions.

This study aims to address the current limitations of dialogue systems by developing an advanced, integrated dialogue system that incorporates multimodal inputs and utilizes generative AI, including LLMs like ChatGPT. By incorporating techniques such as transfer learning, this system will support multilingual interactions and adapt to various user contexts, even in environments with limited data. The objective is to create a multimodal cross-lingual task-oriented dialogue system that enhances the functionality and inclusivity of traditional dialogue systems, thereby improving accessibility and equity in human-computer interactions across diverse user groups. This approach promises to push the boundaries of NLP and set new benchmarks for inclusive technology.

II. RELATED WORKS

In this section, we will review foundational theories and current models that resonate with our objectives, preparing the ground for a detailed discussion on how our proposed system can bridge existing gaps and enhance the capabilities of task-oriented dialogue systems. This approach underscores our commitment to advancing the field of natural language processing by making technology more inclusive and effective for a broad range of user needs.

A. Accessibility and Inclusivity in Dialogue Systems

Domain reasoning is crucial for attaining the required awareness and understanding of user input. This involves interpreting both verbal communication and gestures, and integrating this information with the existing context to form a cohesive understanding of the situation. Such comprehension is essential for the system to take appropriate actions. The recognition text and the given marked text. Jeon & Lee (2022)[21] proposed a multi-domain task-oriented dialogue system using effective context Optimize the dialogue system of the loop action strategy, mainly using supervised learning combined with applied reinforcement learning to optimize the dialogue using the loop dialogue strategy system. This conversational strategy repeatedly generates explicit systematic actions as word response strategies.

B. Evaluation of Language Models in Dialogue Systems

LLMs are developed based on the recognition that enlarging the pre-training corpus and the model size concurrently enhances performance across various NLP tasks [22], Researchers focus on scaling both dimensions to improve sample efficiency, which allows the model to learn more complex patterns and representations from the data. Consequently, PLM systems evolve into LLMs, examples of which include GLM, LLaMA, and InstructGPT [5], [23], [24]. Many Transformer-derived models and pre-training based on the self-attention mechanism Language models, such as BERT ,BART, GPT, GPT-2, GPT-3, ChatGPT, T5, etc., have made great progress in research related to natural language processing[15]. In addition to the field of natural language processing research, Transformer has also been widely used in computer vision in recent years and has achieved great research results.

C. Multimodal Interfaces in Conversational AI

Johnston et al. [25] introduce MATCH, an architecture designed for multimodal dialogue systems that combine speech and graphical input to enhance user interaction in a mobile environment. It addresses critical challenges in interaction design and usability in mobile settings, paving the way for more intuitive and effective human-computer interactions in daily tasks and navigation. Baltrušaitis et al., examine how multimodal learning can be specifically applied to enhance dialogue systems. By integrating visual, textual, and auditory data, dialogue systems can become more context-aware and responsive to the nuances of human communication[26].

D. Cross-Language Capabilities in Task-Oriented Dialogue System

Due to the rapid development of globalization, cross-language dialogue systems are becoming increasingly important in e-commerce and customer service. In real-world cross-language dialogue system deployments, machine translation (MT) services are often used before and after the dialogue system to bridge different languages. However, the noise and errors introduced during machine translation may lead to a reduction in the stability of the dialogue system, making the performance of traditional cross-language dialogue systems still unsatisfactory. Xiang et al. (2021) proposed a machine translation MT noise enhancement framework that utilizes multi-granularity MTnoises and injects such noise into the dialogue system to improve the stability of the dialogue system. Xiang et al. (2021)'s experimental results on three dialogue models, two dialogue datasets, and two language pairs show that the machine translation MT noise enhancement framework significantly improves the performance of the cross-language dialogue system[27].

E. Generative AI for Personalization in Dialogue Systems

Generative AI is a dynamic field that merges technological innovation with creative potential, significantly impacting various sectors while presenting distinct ethical and technical challenges. Unlike traditional systems that pull responses from a set database, generative AI systems use models, typically auto-regressive ones, to create responses word by word[2]. Each new word depends on the words that came before it. The most commonly employed models in these systems are neural generative models, particularly sequence-to-sequence (seq2seq) models. A key benefit of these generative systems is their ability to create responses that are not pre-existing in the training dataset[28].

F. Task-Oriented Dialogue Systems and User Assistance

Subsequently, several other studies used GPT-2 for TOD systems[10][29][30].Hosseini-Asl et al.[10] proposed SimpleTOD, which considered TOD a single-sequence prediction task, and trained a single model with single, joint, multitask loss. They employed transfer learning from pre-trained GPT-2. Peng et al.[30] proposed SOLOIST, which also used GPT-2 and machine teaching for an end-to-end TOD system in both single- and multi-domain settings. Their experiments showed promising performance in a few-shot fine-tuning setting.

This literature review examines the notable advancements in dialogue systems, emphasizing the role of Generative AI and LLMs in advancing natural language processing. It also discusses the enhancement of user interactions through multimodal interfaces, including text and audio, and the development of multilingual TOD systems that support global communication. These systems utilize the latest AI technologies to deliver dynamic, context-aware responses and adapt to diverse linguistic and modal requirements, illustrating the impact of cutting-edge technology on the evolution of conversational AI. However, the review also identifies a gap in the current research landscape, noting a scarcity of studies specifically focusing on the field of inclusive communication, which is crucial for developing truly accessible and equitable dialogue systems.

III. RESEARCH METHODOLOGY AND SYSTEM ARCHITECTURE

A. Research Methodology

In our research, we employed the System Development Research Methodology as outlined in reference [31]. This methodology guided the development of our system through five distinct stages. Each stage is detailed below, describing the specific processes and steps undertaken in our study:

1. Construct a Conceptual Framework:

In this study, we will develop a conceptual framework for the model. We will pinpoint the existing research gaps and conduct a thorough analysis of the requirements necessary for the model system to function optimally.

2. Develop a System Architecture:

Our goal is to develop a scalable and adaptable architecture for an inclusive technology based on multimodal cross-lingual TOD systems using Generative AI. This involves defining the interrelationships and clarifying the functions of various components within the model to ensure it caters effectively to diverse user needs.

3. Analyze and Design the System:

In this research, we aim to design, compile, and develop a comprehensive knowledge base specific to TOD systems using Generative AI. Simultaneously, our objective is to devise the most effective operational procedures in Generative AI in multimodal, cross-lingual TOD systems for inclusive communication systems.

4. Build the System Through the Construction of TOD:

This study aims to explore fundamental concepts, frameworks, and design paradigms to deepen our understanding of the subject.

5. Observe and Evaluate the System:

This research strives to determine the viability of TOD systems and aims to evaluate their performance and accuracy by comparing systematic observations.

B. Proposed system architecture

In our research, we explored a comparative framework that assesses the interplay between LLMs, and multimodal approaches in cross-lingual task-oriented dialogue systems

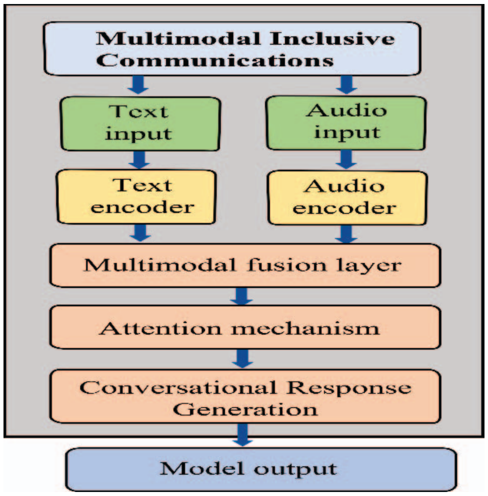

Fig.1 The Proposed System Architecture of Inclusive Communication Multimodal Cross-language Task-Oriented Dialogue System.

using generative Artificial intelligence (AI) for inclusive communication system shown in Fig. 1.

The proposed architecture features multimodal encoders, including text and audio allowing it to process diverse types of input. Users can submit their requirements in text formats. After encoding the data, the system employs a fusion layer and an attention mechanism to retrieve relevant information. This enables the system to generate conversational responses and complete specific tasks effectively. The underlying pre-trained foundation model that supports these capabilities will be described in detail below,

In the text, for the encoder-decoder architecture and cross-lingual feature of the model, we use the GPT 3.5 turbo for generating text and we use Llama 2 [32]to be the text encoder and decoder. And. At last, for the audio input, we made use of Whisper [33] as our audio encoder and decoder.

C. Data collection

To train our system's model, we required a dataset that includes multilingual data. During the data collection phase of our research, we integrated the generated dataset generated from GPT 3.5 turbo (Fig. 2) to construct a new multimodal and multilingual dataset tailored to our needs due to data scarcity. We utilized healthcare data about Amyotrophic Lateral Sclerosis (ALS) in our study. The data was collected from the Taiwan motor neuron disease association and generated into the Question and answering dataset.

D. Evaluation metrics

To benchmark our experimental model against prior research, we will evaluate it using Accuracy, Mean Average Precision (mAP), and Precision (Prec) as our key metrics, consistent with those used in [34]. These metrics range from 0 to 1, where higher scores indicate better performance, particularly for mAP and Precision. The formulas for mAP and Precision are detailed below,

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}, \quad (1)$$

$$\text{Precision} = \frac{TP}{TP+FP}, \quad (2)$$

$$\text{mean Average Precision} = \frac{1}{N}\sum_{i=1}^{N} AP_i, \quad (3)$$

{"role": "system", "content": "你是一個語言障礙疾病與照護的專家."},
{"role": "user", "content": "1. ALS是什麼疾病?它的病徵有哪些?"},
{"role": "assistant", "content": "肌萎縮性側索硬化症(ALS)是一種運動神經元進行性退化的疾病,導致肌肉無法得到正常刺激而逐漸萎縮和無力。常見病徵有肌肉無力、肌肉萎縮、肌肉痙攣、吞嚥困難、言語不清等。"},
{"role": "user", "content": "2. ALS患者可能出現的呼吸困難症狀有哪些?需要怎麼處理?"},
{"role": "assistant", "content": "ALS患者可能出現的呼吸困難症狀包括呼吸肌無力導致呼吸不順、呼吸時胸腹不協調、打鼾、睡眠時呼吸中斷等。建議使用呼吸器輔助通氣,並進行換氣和拍痰物理治療,必要時可考慮進行氣切開孔手術改善呼吸。"},
{
"Q": "What is ALS?",
"A": "Amyotrophic Lateral Sclerosis (ALS) is a progressive neurodegenerative disease that affects nerve cells in the brain and spinal cord, leading to the loss of voluntary muscle movement."
},
{
"Q": "How is ALS different from other motor neuron diseases?",
"A": "ALS is a type of motor neuron disease characterized by the degeneration of both upper and lower motor neurons. It is the most common form of motor neuron disease."

Fig. 2 Generated text sample used in this study

IV. DATA ANALYSIS AND DISCUSSION

To construct the architecture of our proposed inclusive technology multimodal cross-lingual task-oriented dialogue system, we conducted a comprehensive analysis using trend analysis, keyword analysis, region analysis, and document analysis of research on TOD systems sourced from Scopus. The findings from these analyses are detailed in this section, providing insights into the current trends, key themes, regional focuses, and significant literature in the field. This analysis forms the basis for understanding the landscape of TOD system research and informs the development of an inclusive dialogue system

A. Trend Analysis

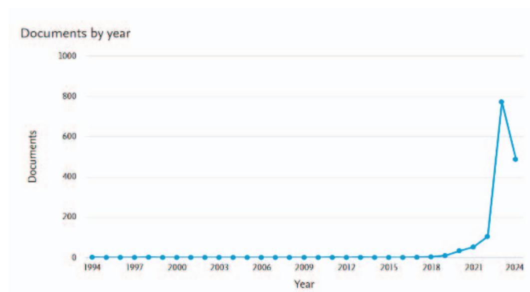

Fig. 3 Trend analysis of task-oriented dialogue system.

Fig. 3 sets out the trend analysis of the relevant papers on task-oriented dialogue systems from 2000 to 2023. We adopted Scopus as our database and used the keywords, including "Generative AI", "dialogue system","Large

Language Models" and "Multimodal" ,"Cross-lingual" in this analysis. What could be seen in Figure 2 was a sustained growth in the amount of research over the past two decades since 2002.

B. Keyword Analysis

Table 1 provides an overview of the research on generative AI that incorporates multimodal, LLMs, and crosslingual based on data from Scopus spanning 2000 to 2023. It reveals that only 3 documents explore multimodal multilingual task-oriented dialogue systems keywords in field of NLP and only 2 documents focus on explore generative AI in multimodal multilingual task-oriented dialogue systems keywords in the field of NLP. Combining multimodal and multilingual architectures in task-oriented dialogue systems is a promising area for further NLP research. Table 2 shows a co-occurrence keyword analysis of Generative AI, covering 1465 keywords from 945 studies, with the top 10 keywords highlighted. Figure 4 displays the top 10 countries contributing to research on task-oriented dialogue systems, according to Scopus data. China and the United States lead with the highest number of publications.

C. Region Analysis

TABLE 1. Keyword Statistics for Multimodal and Multilingual Task-Oriented Dialogue Systems Research in Scopus, 2000-2023

Searching Keywords	Count
"Generative AI " AND "Generative Artificial Intelligence" AND "Generative Ai"	1465
"Generative AI " AND "Generative Artificial Intelligence"	945
"Generative AI " AND "Generative Artificial Intelligence" AND "Generative Ai" AND "Large Language Model" AND Large Language Models" AND"LLM"	30
"Large Language Model" AND "task-oriented dialogue system" AND" multimodal" AND "Cross-lingual "	3
"Generative AI" AND "task-oriented dialogue system" AND" multimodal" AND "Cross-lingual "	2

TABLE 2. Co-Occurrence Keyword of Generative AI in multimodal task oriented system

Rank	Keyword	Occurrence
1.	Generative AI	879
2	Artificial Intelligence	437
3.	ChatGPT	333
4.	Generative Adversarial Networks	178
5.	Language Model	170
6.	Generative Artificial Intelligence	164
7.	Deep Learning	159
8.	Machine Learning	146
9.	Large Language Model	143
10.	Large Language Models	141

Table 3 details the leading affiliations in task-oriented dialogue system research, based on Scopus data. Notably, the top six affiliations originate from the two countries previously identified as the most prolific in this research area.

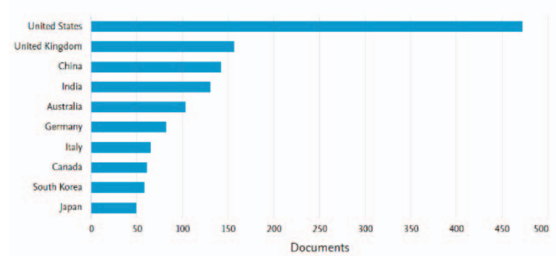

Fig. 4 The Top 10 Country from The Research of Task-Oriented Dialogue System in Scopus

TABLE 3. The Top 10 Affiliation from The Research of Task-Oriented Dialogue System in Scopus

Rank	Country	Count
1.	Nanyang Technological University	19
2.	The Hong Kong Polytechnic University	17
3.	Monash University	16
4.	Carnegie Mellon University	15
5.	IBM Research	14
6.	Massachusetts Institute of Technology	14
7.	Columbia University	13
8.	National University of Singapore	12
9.	Queen Mary University of London	12
10.	UNSW Sydney	12

V. CONCLUSION

Innovative AI systems with multimodal structures and generative AI technologies enhance dialogue interactions and data management. They improve task completion capabilities, even with limited data, and boost efficiency through cross-lingual transfer learning. By enabling inclusive communication, these systems transcend language barriers, serving a broader user base and paving the way for versatile AI applications. Generative AI-driven task-oriented dialogue systems represent a significant advancement, supporting multilingual, multimodal interactions, breaking language barriers, and providing tailored responses, thereby fostering digital inclusion and equitable human-computer interactions.

The major research contribution of this study is the potential of generative AI in developing accessible dialogue systems that cater to diverse user needs to advance inclusive technology. This research presents an overview of current advancements in the field, highlighting the promise of integrated dialogue systems leveraging multimodal inputs and LLMs.

Practitioner implications highlight the system's role in fostering digital inclusion and communication support, enhancing accessibility and equity for diverse users.

Research limitations include constraints due to the quality and diversity of training data, affecting generalizability across languages and modalities, high computational requirements, and limited accessibility. Further research is needed to overcome these challenges.

ACKNOWLEDGMENTS

This research was supported in part by the National Science and Technology Council (NSTC), Taiwan, under grants NSTC 112-2425-H-305-002-, and NSTC 112-2627-M-038-001-, and National Taipei University (NTPU), Taiwan under grants 113-NTPU_ORDA-F-003, 113-NTPU-ORDA-F-004, USTP-NTPU-TMU-113-03, NTPU-112A413E01, and NTPU-113A513E01.

REFERENCES

[1] A. Papangelis, G. Galatas, and K. Tsiakas, "LNCS 8515 - A Dialogue System for Ensuring Safe Rehabilitation," pp. 349–358, 2014.

[2] M. Huang, X. Zhu, and J. Gao, "Challenges in Building Intelligent Open-domain Dialog Systems," *ACM Trans. Inf. Syst.*, vol. 38, no. 3, 2020, doi: 10.1145/3383123.

[3] Z. Zhang, R. Takanobu, Q. Zhu, M. L. Huang, and X. Y. Zhu, "Recent advances and challenges in task-oriented dialog systems," *Sci. China Technol. Sci.*, vol. 63, no. 10, pp. 2011–2027, 2020, doi: 10.1007/s11431-020-1692-3.

[4] L. Zhou, J. Gao, D. Li, and H. Y. Shum, "The design and implementation of xiaoice, an empathetic social chatbot," *Comput. Linguist.*, vol. 46, no. 1, pp. 53–93, 2020, doi: 10.1162/COLI_a_00368.

[5] L. Ouyang *et al.*, "Training language models to follow instructions with human feedback," *Adv. Neural Inf. Process. Syst.*, vol. 35, no. NeurIPS, 2022.

[6] H. Wang *et al.*, "Cue-CoT: Chain-of-thought Prompting for Responding to In-depth Dialogue Questions with LLMs," *Find. Assoc. Comput. Linguist. EMNLP 2023*, pp. 12047–12064, 2023, doi: 10.18653/v1/2023.findings-emnlp.806.

[7] Y. Huang, X. Feng, X. Feng, and B. Qin, "The Factual Inconsistency Problem in Abstractive Text Summarization: A Survey," 2021, [Online]. Available: http://arxiv.org/abs/2104.14839

[8] A. Salemi, S. Mysore, M. Bendersky, and H. Zamani, "LaMP: When Large Language Models Meet Personalization," 2023, [Online]. Available: http://arxiv.org/abs/2304.11406

[9] Z. Liu, J. Xu, Z. Lei, H. Wang, Z. Y. Niu, and H. Wu, "Where to Go for the Holidays: Towards Mixed-Type Dialogs for Clarification of User Goals," *Proc. Annu. Meet. Assoc. Comput. Linguist.*, vol. 1, pp. 1024–1034, 2022, doi: 10.18653/v1/2022.acl-long.73.

[10] E. Hosseini-Asl, B. McCann, C. S. Wu, S. Yavuz, and R. Socher, "A simple language model for task-oriented dialogue," *Adv. Neural Inf. Process. Syst.*, vol. 2020-Decem, no. NeurIPS, 2020.

[11] H. Fausk and D. C. Isaksen, "Improving Language Understanding by Generative Pre-Training," *Homol. Homotopy Appl.*, vol. 9, no. 1, pp. 399–438, 2007.

[12] V. K. Tran, L. M. Nguyen, and S. Tojo, "Neural-based natural language generation in dialogue using RNN encoder-decoder with semantic aggregation," *SIGDIAL 2017 - 18th Annu. Meet. Spec. Interes. Gr. Discourse Dialogue, Proc. Conf.*, pp. 231–240, 2017, doi: 10.18653/v1/w17-5528.

[13] F. Petroni *et al.*, "Language models as knowledge bases?," *EMNLP-IJCNLP 2019 - 2019 Conf. Empir. Methods Nat. Lang. Process. 9th Int. Jt. Conf. Nat. Lang. Process. Proc. Conf.*, pp. 2463–2473, 2019, doi: 10.18653/v1/d19-1250.

[14] Y. Zhang *et al.*, "DIALOGPT: Large-scale generative pre-training for conversational response generation," *Proc. Annu. Meet. Assoc. Comput. Linguist.*, pp. 270–278, 2020, doi: 10.18653/v1/2020.acl-demos.30.

[15] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, no. Mlm, pp. 4171–4186, 2019.

[16] A. Conneau *et al.*, "Unsupervised cross-lingual representation learning at scale," *Proc. Annu. Meet. Assoc. Comput. Linguist.*, pp. 8440–8451, 2020, doi: 10.18653/v1/2020.acl-main.747.

[17] L. Xue *et al.*, "mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer," *NAACL-HLT 2021 - 2021 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. Proc. Conf.*, pp. 483–498, 2021, doi: 10.18653/v1/2021.naacl-main.41.

[18] S. Wu and M. Dredze, "Beto, Bentz, Becas: The surprising cross-lingual effectiveness of Bert," *EMNLP-IJCNLP 2019 - 2019 Conf. Empir. Methods Nat. Lang. Process. 9th Int. Jt. Conf. Nat. Lang. Process. Proc. Conf.*, pp. 833–844, 2019, doi: 10.18653/v1/d19-1077.

[19] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 1, pp. 346–350, 2021, doi: 10.21437/Interspeech.2021-329.

[20] A. Ansell *et al.*, "MAD-G: Multilingual Adapter Generation for Efficient Cross-Lingual Transfer," *Find. Assoc. Comput. Linguist. Find. ACL EMNLP 2021*, vol. 000, pp. 4762–4781, 2021, doi: 10.18653/v1/2021.findings-emnlp.410.

[21] H. Jeon and G. G. Lee, "DORA: Towards policy optimization for task-oriented dialogue system with efficient context," *Comput. Speech Lang.*, vol. 72, no. April 2021, p. 101310, 2022, doi: 10.1016/j.csl.2021.101310.

[22] J. Kaplan *et al.*, "Scaling Laws for Neural Language Models," 2020, [Online]. Available: http://arxiv.org/abs/2001.08361

[23] Z. Du *et al.*, "GLM: General Language Model Pretraining with Autoregressive Blank Infilling," *Proc. Annu. Meet. Assoc. Comput. Linguist.*, vol. 1, pp. 320–335, 2022, doi: 10.18653/v1/2022.acl-long.26.

[24] H. Touvron *et al.*, "LLaMA: Open and Efficient Foundation Language Models," 2023, [Online]. Available: http://arxiv.org/abs/2302.13971

[25] M. Johnston *et al.*, "MATCH: An Architecture for Multimodal Dialogue Systems," *Proc. Annu. Meet. Assoc. Comput. Linguist.*, vol. 2002-July, no. July, pp. 376–383, 2002.

[26] T. Baltrusaitis, C. Ahuja, and L. P. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, 2019, doi: 10.1109/TPAMI.2018.2798607.

[27] L. Xiang, J. Zhu, Y. Zhao, Y. Zhou, and C. Zong, "Robust Cross-lingual Task-oriented Dialogue," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 20, no. 6, 2021, doi: 10.1145/3457571.

[28] H. Chen, X. Liu, D. Yin, and J. Tang, "A Survey on Dialogue Systems," *ACM SIGKDD Explor. Newsl.*, vol. 19, no. 2, pp. 25–35, 2017, doi: 10.1145/3166054.3166058.

[29] J. Kulhánek, V. Hudeček, T. Nekvinda, and O. Dušek, "AuGPT: Auxiliary Tasks and Data Augmentation for End-To-End Dialogue with Pre-Trained Language Models," *NLP Conversational AI, NLP4ConvAI 2021 - Proc. 3rd Work.*, pp. 198–210, 2021, doi: 10.18653/v1/2021.nlp4convai-1.19.

[30] B. Peng, C. Li, J. Li, S. Shayandeh, L. Liden, and J. Gao, "Soloist: Building task bots at scale with transfer learning and machine teaching," *Trans. Assoc. Comput. Linguist.*, vol. 9, pp. 807–824, 2021, doi: 10.1162/tacl_a_00399.

[31] J. F. Nunamaker, M. Chen, and T. D. M. Purdin, "Systems development in information systems research," *J. Manag. Inf. Syst.*, vol. 7, no. 3, pp. 89–106, 1990, doi: 10.1080/07421222.1990.11517898.

[32] H. Touvron *et al.*, "Llama 2: Open Foundation and Fine-Tuned Chat Models," 2023, [Online]. Available: http://arxiv.org/abs/2307.09288

[33] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," *Proc. Mach. Learn. Res.*, vol. 202, pp. 28492–28518, 2023.

[34] M. Pretraining *et al.*, "M5Product : Self-harmonized Contrastive Learning for E-commercial," *Cvpr*, pp. 21252–21262, 2022.