A Survey of Large Language Models and Their Role in Task-Oriented Dialogue Systems

Abstract

Large Language Models (LLMs) have revolutionized Natural Language Processing (NLP) and Task-Oriented Dialogue Systems (TDSs), significantly enhancing language comprehension, generation, and dialogue management. This survey explores the integration of LLMs in TDSs, highlighting their role in improving conversational AI through advanced capabilities in context management and response generation. The survey underscores LLMs' transformative impact across domains, including healthcare and e-commerce, by facilitating personalized interactions and enhancing user satisfaction. Despite their advancements, LLMs face challenges such as ethical concerns, computational constraints, and data biases, necessitating ongoing research to optimize their deployment and ensure responsible AI use. The integration of knowledge graphs and retrieval-augmented generation techniques into LLMs enhances factual accuracy and context relevance, addressing issues like hallucination and inconsistency. Future research should focus on improving LLM efficiency, addressing biases, and exploring domain-specific applications to maximize their potential. This survey emphasizes the need for comprehensive evaluation frameworks and ethical guidelines to guide LLM development, ensuring their continued evolution and effectiveness in diverse applications.

1 Introduction

1.1 Overview of Large Language Models (LLMs)

Large Language Models (LLMs) have become pivotal in Natural Language Processing (NLP), enhancing language comprehension and generation across various applications, including machine translation, question-answering, and keyword extraction. Utilizing advanced architectures such as Generative Pre-trained Transformers (GPT) and Bidirectional Encoder Representations from Transformers (BERT), LLMs leverage extensive datasets to discern intricate linguistic patterns and semantic nuances, proving essential for tasks like sentiment analysis and text summarization [1, 2].

The evolution of LLMs from static knowledge bases to dynamic agents capable of real-world actions marks a significant advancement in AI, particularly in task-oriented dialogue systems, where they facilitate natural language interactions that enhance task completion. However, challenges remain, especially in achieving sophisticated reasoning abilities for complex tasks [3].

Moreover, LLMs bridge linguistic divides in multilingual and multimodal contexts, adeptly processing semi-structured data from diverse sources, such as PDFs, and converting it into structured formats [4]. The rapid evolution of these models has also raised concerns regarding the differentiation between human and LLM-generated content, highlighting the need for advanced detection methodologies [5].

In the medical field, LLMs necessitate tailored evaluation frameworks to ensure ethical implementation [6]. Their intersection with security and privacy issues presents both opportunities and risks, requiring a thorough understanding of vulnerabilities to guide future research. LLMs enhance context understanding and relationship extraction within texts, crucial for tasks like slot filling and intent classification in dialogue systems.

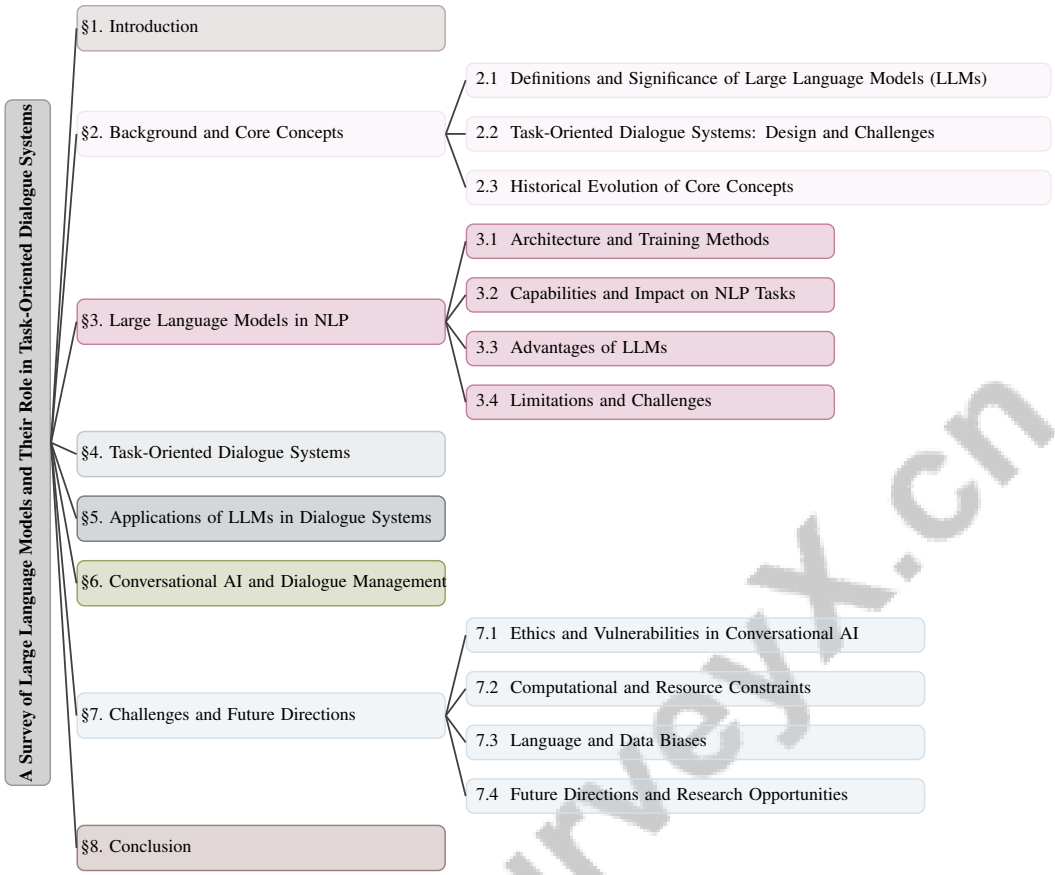


Figure 1: chapter structure

As LLMs evolve, they promise to unlock new opportunities in AI-driven language technologies, addressing challenges such as social biases and toxic content, while improving inference control and explainability [7]. Their future hinges on navigating these complexities and continuing to innovate within this dynamic landscape.

1.2 Interconnectedness with Task-Oriented Dialogue Systems

The integration of LLMs into Task-Oriented Dialogue Systems (TDSs) signifies a major leap in conversational AI, enhancing the management of complex interactions and user satisfaction. Traditional TDSs, characterized by modular frameworks for dialogue state tracking, management, and natural language understanding, often face inefficiencies and high annotation costs. In contrast, LLMs streamline these processes through end-to-end learning, reducing reliance on annotated datasets and improving dialogue management efficiency [8].

LLMs excel in handling complex compositional semantics, which traditional intent-slot frameworks struggle to address. This capability fosters a nuanced understanding of user intents, enabling the decomposition of these intents into executable actions for accurate request fulfillment. Furthermore, LLMs model long dialogue contexts and incorporate external knowledge, essential for maintaining coherence in extended conversations [9].

Despite their advantages, deploying LLMs in TDSs presents challenges, including security, privacy, and ethical concerns that necessitate robust control frameworks to ensure reliability and efficiency [10]. Additionally, limited domain API coverage in many TDSs can lead to inefficiencies when users request information beyond available APIs [11].

LLMs also automate traditionally labor-intensive data processing tasks, enhancing systems' capabilities to manage diverse and complex queries efficiently [4]. Their application in generating theme-aware keywords for E-commerce products exemplifies their versatility across domains [2].

Incorporating knowledge graphs (KGs) can mitigate issues like hallucination and inconsistency in LLM outputs, emphasizing the interconnected nature of LLMs and TDSs [9].

As LLMs advance, their role in enhancing TDSs will be critical, offering sophisticated and context-aware conversational AI solutions. Their integration promises to address existing challenges, such as the need for continuous learning mechanisms to prevent catastrophic forgetting and ensure knowledge retention over time [6].

1.3 Scope and Objectives of the Survey

This survey provides a comprehensive exploration of LLMs and their integration into TDSs, emphasizing implications for Conversational AI and Dialogue Management. The scope encompasses an examination of LLM fundamentals, architectural innovations, training strategies, and applications across diverse domains such as telecommunications, medicine, education, finance, and engineering [12]. A critical objective is to assess ethical concerns related to LLMs, including privacy, fairness, misinformation, accountability, and governance, which are increasingly relevant as these models gain practical traction [13].

The survey evaluates challenges in LLM development, focusing on pre-training, adaptation tuning, and utilization, while excluding unrelated topics [14]. Methodologically, it covers recent advancements in LLM training paradigms, including pre-training followed by fine-tuning, prompt-based learning, and the reformulation of NLP tasks as text generation [15]. Additionally, it provides an overview of evaluation methods for TDSs, emphasizing practical applications, evaluation challenges, and proposing future research directions [16].

Furthermore, the survey investigates LLM versatility in academic contexts, examining their applications in writing, coding, data analysis, and literature review, highlighting both potential and implementation challenges [17]. It also explores tool learning methodologies organized into stages such as task planning, tool selection, tool calling, and response generation, underscoring the advantages of tool integration [18]. By encompassing a wide range of applications, from LLM-empowered recommender systems to machine learning in healthcare, this survey aims to present a holistic view of the current landscape and future directions in LLM integration within dialogue systems [19].

The objectives are to synthesize existing research, identify key challenges, and propose future research directions to advance LLMs and their applications in TDSs, while addressing open challenges across various industries [7].

1.4 Structure of the Survey

This survey is structured to provide a thorough examination of LLMs and their integration within TDSs, while exploring broader implications in Conversational AI and Dialogue Management. The introductory section establishes a foundation by presenting an overview of LLMs, detailing their significance in NLP, and discussing their interconnectedness with TDSs. It outlines the survey's scope and objectives, setting the stage for subsequent analyses.

Following the introduction, the survey delves into background and core concepts, providing definitions and discussing the historical evolution of LLMs, NLP, and dialogue systems. This section equips readers with contextual understanding of the technological advancements leading to the current landscape.

The survey then examines LLM roles in NLP, addressing architecture, training methodologies, capabilities, and impacts on various NLP tasks. It presents a balanced view of advantages, limitations, and challenges associated with LLMs.

Subsequently, the paper explores LLM integration in TDSs, analyzing how LLMs enhance intent detection, user interaction, and facilitate complex tasks such as speech summarization and academic writing assistance [20, 21, 22, 23].

The applications of LLMs in dialogue systems are further examined, highlighting innovative applications, domain-specific enhancements, and multilingual capabilities. This section underscores the adaptability of LLMs across sectors and languages.

The survey delves into Conversational AI and Dialogue Management, specifically how LLMs like ChatGPT enhance conversation flow and context management, transforming user interactions by leveraging vast data for improved response generation and knowledge creation [22, 21, 7]. It also investigates advanced dialogue management techniques, emphasizing LLM roles in refining these processes.

The penultimate section identifies key challenges and proposes future directions for LLM integration in dialogue systems, addressing ethical considerations, computational constraints, and biases, while suggesting research opportunities to overcome these challenges.

Finally, the conclusion synthesizes key findings and insights, emphasizing the importance of ongoing research and innovation in LLMs and their dialogue system applications. This structured approach facilitates comprehensive examination and yields significant insights for researchers and practitioners by leveraging methodologies such as Theme-Aware Keyword Extraction with LLMs and AI-assisted data processing techniques, enhancing the accuracy and efficiency of keyword generation and data management processes [4, 2].The following sections are organized as shown in Figure 1.

2 Background and Core Concepts

2.1 Definitions and Significance of Large Language Models (LLMs)

Large Language Models (LLMs), including GPT and BERT, signify a pivotal advancement in AI by enabling sophisticated comprehension, generation, and manipulation of human language. These models have revolutionized Natural Language Processing (NLP) through deep learning, analyzing vast text corpora to uncover complex linguistic and semantic patterns, achieving leading performance across numerous NLP tasks often without extensive task-specific training [24]. Beyond traditional applications, LLMs are increasingly applied in specialized domains like healthcare for tasks such as sentiment analysis and dialogue management, although English-centric datasets pose challenges [13]. Their ability to interpret and execute natural language instructions is crucial for developing AI agents capable of strategic negotiation and long-term planning [9]. Retrieval-Augmented Generation (RAG) methods further enhance LLMs' capacity to produce accurate, contextually relevant responses by integrating information retrieval into text generation [9].

LLMs' integration into research workflows, particularly in competitive markets, highlights their potential to enhance productivity and creativity, despite challenges related to symbol grounding and ethical considerations [25]. The opaque nature of LLM operations complicates transparency, affecting stakeholders' understanding of their capabilities and limitations [13]. Additionally, LLMs face difficulties in recalling relevant knowledge and generating knowledge-grounded content, limiting their effectiveness in tasks requiring factual reasoning [9]. In security contexts, while enhancing measures, LLMs also introduce new vulnerabilities [13]. The challenge of distinguishing LLM-generated text from human-authored content necessitates advanced detection methods, as seen in initiatives like GigaCheck [5]. Lifelong learning strategies are crucial for LLMs to adapt to new knowledge without altering existing parameters [9].

2.2 Task-Oriented Dialogue Systems: Design and Challenges

Task-Oriented Dialogue Systems (TDSs) aim to facilitate structured interactions to achieve specific user goals, employing a modular architecture comprising components like Natural Language Understanding (NLU), Dialogue State Tracking (DST), Dialogue Policy (DP), and Natural Language Generation (NLG). While this modular approach aids task delineation, it presents integration challenges, especially in managing multilingual content and varied data formats [4]. Generating user-request-fulfilling utterances is complicated by the need to incorporate complex linguistic and contextual information [26]. High costs and complexities of obtaining high-quality annotated dialogue data hinder robust DST model development [27]. The interplay between Slot Filling (SF) and Intent Classification (IC), along with the demand for large labeled datasets, complicates model adaptation to new domains, particularly when data is scarce [28]. Existing user simulation models often fail to produce diverse and realistic interactions, limiting their effectiveness in evaluating and enhancing dialogue systems [29].

Incorporating external knowledge into dialogue systems remains a critical challenge, with obstacles in knowledge-seeking turn detection, selection, and generation of knowledge-grounded responses [11].

Increasing complexity in maintaining dialogue logic in rule-based systems adds further difficulty [30]. Although LLMs have improved user interactions, they struggle to maintain current knowledge and address logical reasoning deficiencies, leading to issues like hallucination and unreliable outputs. Current evaluation benchmarks focus mainly on binary classification, inadequately capturing the nuances of Human-Machine collaborative texts, thus limiting practical TDS assessment [5]. The limited adaptive learning capabilities of LLMs restrict their effectiveness in replicating nuanced and responsive behaviors required in dynamic environments [25].

2.3 Historical Evolution of Core Concepts

The historical evolution of NLP, LLMs, and dialogue systems is marked by significant advancements shaping the current AI landscape. Language technologies began with rule-based systems, establishing foundational principles for subsequent advancements in natural language understanding (NLU) and generation (NLG). These early systems paved the way for pre-trained language models (PLMs) and knowledge-enhanced PLMs (KE-PLMs), leveraging extensive text corpora and external knowledge to overcome reasoning and comprehension limitations. Consequently, models like ChatGPT have emerged, significantly influencing academic contexts and challenging traditional skill acquisition and assessment methods [22, 31].

Initial NLP phases focused on syntactic parsing and basic language understanding, crucial for developing early dialogue systems reliant on manually crafted rules and templates, limiting flexibility and scalability [32]. The introduction of Statistical Language Models (SLMs) marked a pivotal shift, enabling robust language processing capabilities through probabilistic approaches [14]. The advent of Neural Language Models (NLMs) revolutionized NLP by utilizing deep learning techniques to enhance language task accuracy and efficiency. This era saw the emergence of Pre-trained Language Models (PLMs) that utilized large-scale datasets to learn general language representations, significantly improving downstream application performance [14]. The transition to LLMs like GPT and BERT exemplified these models' scaling effects and emergent abilities, leading to unprecedented advancements in language comprehension and generation [14].

Simultaneously, task-oriented dialogue (TOD) systems evolved from simple rule-based interactions to sophisticated frameworks incorporating NLU, DST, and Response Generation (RG), refined to support more complex and dynamic conversational scenarios [33]. Datasets like MultiWOZ, collected through a Wizard-of-Oz setup, provided a large-scale, multi-domain corpus instrumental in advancing dialogue system research [34]. Despite advancements, TOD system development has predominantly focused on high-resourced languages, resulting in performance disparities in multilingual contexts, highlighting the need for inclusive approaches addressing challenges faced by low-resourced languages [35]. The Dialog-To-Actions (DTA) method, representing natural language responses as sequences of dialogue actions, exemplifies innovations enhancing dialogue system reliability and efficiency compared to traditional token-level generation methods [36].

The historical trajectory of NLP, LLMs, and dialogue systems reflects continuous innovation driven by advancements in computational methodologies and diverse dataset availability. Progress in LLMs and AI has significantly shaped sophisticated conversational AI systems, revealing remarkable capabilities in natural language processing alongside complex challenges, including ethical concerns, bias, and responsible usage guidelines. These developments emphasize ongoing research and collaboration among stakeholders to address AI implications across various sectors, maximizing benefits while minimizing associated risks [37, 38, 39, 22, 40].

3 Large Language Models in NLP

Large Language Models (LLMs) have become indispensable in Natural Language Processing (NLP), enhancing tasks such as machine translation, question-answering, and content generation. These models utilize sophisticated architectures and training techniques to discern intricate linguistic patterns, producing coherent and contextually appropriate outputs. Their application spans sectors like e-commerce, education, and social networks, aiding in tasks such as keyword extraction and content moderation. Despite their benefits, LLMs present ethical dilemmas, biases, and challenges in interpretability that researchers are striving to resolve to improve their reliability in real-world scenarios [12, 47, 2, 1, 48]. Table 1 presents a detailed summary of the methods and features relevant to Large Language Models (LLMs), showcasing their architectural innovations, capabilities, and

Category	Feature	Method
Architecture and Training Methods	Unsupervised Learning Approaches Knowledge and Module Integration Dialogue and Interaction Management Task-Specific Adaptation	RBRM[41] GPT-ACN[42] CALM[8] MNLG[43]
Capabilities and Impact on NLP Tasks	Generalization and Adaptability Multimodal and Integration Interactive Learning	AIDP[4], LMA[44] RAG[24], MMLLMAIS[45] UBAR[46]
Advantages of LLMs	Comprehension Improvement	HDNO[26], LUAS[27]
Limitations and Challenges	Thematic Contextualization	LLM-TAKE[2]

Table 1: This table provides a comprehensive overview of various methodologies and features associated with Large Language Models (LLMs) in Natural Language Processing (NLP). It categorizes the methods based on their architecture, training strategies, capabilities, advantages, and limitations, highlighting key approaches and their contributions to advancing NLP tasks.

the challenges they face in the field of Natural Language Processing (NLP). Additionally, Table 3 provides a comprehensive comparison of different methods employed by Large Language Models (LLMs), detailing their training paradigms, application domains, and the challenges they encounter in Natural Language Processing (NLP). An in-depth exploration of LLMs requires understanding their architecture and training methods, which are crucial to their performance in complex language tasks.

3.1 Architecture and Training Methods

LLMs' architecture and training methods have evolved significantly, driven by the need for enhanced adaptability and performance in diverse NLP tasks. Transformer-based models like GPT (Generative Pre-trained Transformer) and BERT (Bidirectional Encoder Representations from Transformers) utilize self-attention mechanisms for precise language processing and generation [24]. These models follow a pre-train and fine-tune paradigm, involving extensive unsupervised pre-training followed by task-specific fine-tuning, achieving state-of-the-art results [14].

Innovations such as prompt-based learning reformulate NLP tasks as text generation problems, allowing effective use of pre-trained knowledge without extensive retraining [3]. Frameworks like KB-adapters enhance LLMs by integrating domain-specific knowledge, improving their capability in specialized tasks [42].

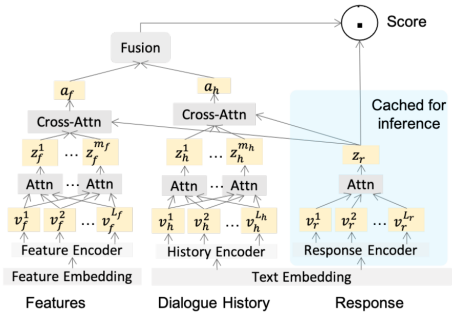
In task-oriented dialogue systems, models like UBAR fine-tune GPT-2 to generate belief states, system acts, and responses based on dialog context, enhancing coherence and relevance [46]. Frameworks such as CALM interpret user messages into executable commands, showcasing LLMs' practical applications in dialogue management [8]. Tailored training approaches are necessary for domain-specific customization, as evidenced by ChatFlow and AutoFlow, which facilitate knowledge transfer and optimize LLM agents through reinforcement learning [49].

Knowledge graph (KG) integration into LLMs through various frameworks enhances model performance, improving response accuracy and relevance [9]. The GigaCheck framework, utilizing the Mistral-7B model, exemplifies advancements in detection methodologies, achieving state-of-the-art results across datasets [5].

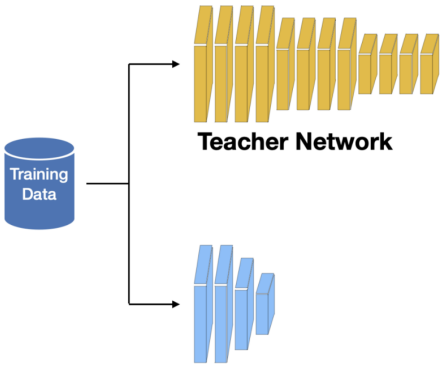
Meta-learning approaches like Meta-NLG optimize natural language generation for low-resource tasks by simulating multiple meta tasks during training, facilitating swift adaptation to new domains [43]. Hierarchical frameworks such as HDNO provide structured dialogue policy and natural language generation by operating at different levels of abstraction [26].

The ongoing evolution of LLM architecture and training methodologies underscores their transformative potential in conversational AI and dialogue management systems, enhancing user interaction and facilitating rapid development of specialized chatbots across various domains, including healthcare and education. Comprehensive evaluation mechanisms comparing LLM performance with human assessments are essential for refining these models and ensuring they meet user needs while addressing data usage and content generation challenges [5, 21, 7, 2, 22].

As illustrated in Figure 2, notable examples in exploring LLM architecture and training methods include "Cross-Attention-based Dialogue Answering with History-Cached Attention" and "Teacher Network." The former utilizes cross-attention mechanisms, structured around feature, history, and response encoders to enhance dialogue interpretation and response generation. The latter demonstrates



(a) Cross-Attention-based Dialogue Answering with History-Cached Attention[41]



(b) Teacher Network[50]

Figure 2: Examples of Architecture and Training Methods

how training data shapes a network's capacity to generalize and perform complex language tasks through multiple layers. These examples reveal the intricate architectures and training methodologies underpinning LLM capabilities in NLP.

3.2 Capabilities and Impact on NLP Tasks

Method Name	Model Capabilities	Training Methodologies	Ethical Concerns
RAG[24]	Text Generation	Fine-tuning	Ethical Implications
MMLLMAIS[45]	Automated Dental Diagnosis	Multimodal Models	Data Privacy Concerns
LMA[44]	Text Generation	Supervised Learning	Ethical Implications
AIDP[4]	Multilingual Content Handling	Hybrid Ai-assisted Approach	Biases IN AI
UBAR[46]	Response Generation	Session-level Training	Computational Demands
LLM-TAKE[2]	Text Generation	Multimodal Models	Ethical Implications

Table 2: Comparison of various methods in large language models (LLMs), highlighting their model capabilities, training methodologies, and associated ethical concerns. This table provides insights into the diverse applications and challenges faced by LLMs in natural language processing (NLP) tasks.

LLMs have transformed the NLP landscape by enhancing performance and versatility in tasks such as text generation, machine translation, and advanced reasoning. However, they also introduce challenges related to ethical concerns, model biases, and computational resource demands that must be addressed for effective deployment in real-world applications [12, 1]. The evolution of LLMs is characterized by advancements in supervised learning, unsupervised pre-training, fine-tuning, and the emergence of multimodal models, enabling architectures like GPT, BERT, and T5 to leverage extensive datasets for learning intricate linguistic patterns and semantic relationships. Figure 3 illustrates these capabilities and challenges, highlighting enhancements in text generation, machine translation, and reasoning, alongside ethical concerns and resource demands. Additionally, Table 2 presents a detailed comparison of different LLM methods, illustrating their capabilities, training approaches, and ethical considerations within the context of NLP advancements.

In text summarization, LLMs outperform traditional models, achieving superior BLEU, ROUGE, and BERT scores [24]. Methodologies such as black-box optimization for instruction generation further enhance LLM adaptability, enabling dynamic proposal and refinement of instructions [51]. Notably, LLMs excel in zero-shot settings, demonstrating competitive performance without task-specific training, highlighting their potential in reasoning tasks [51].

In specialized domains, LLMs automate diagnosis and treatment planning in dentistry, improving clinical workflow efficiency and accuracy [45]. In education, their lifecycle involves development and customization phases, tailoring models for specific tasks and demonstrating their impact on educational NLP applications [52]. Prompt utilization to guide LLMs illustrates their capabilities in stylistic and creative language tasks [51].

Despite their strengths, LLMs face challenges in summarization and sequence tagging, where limitations have been observed [44]. Ethical concerns and biases inherent in LLMs necessitate ongoing scrutiny to ensure responsible deployment and future advancements [4].

In conversational systems, reinforcement learning (RL) optimizes control policies and adapts to user interactions, leading to improved performance [46]. Innovative frameworks enhance model capacity and efficiency in dialogue systems, balancing performance with computational cost [53]. The combination of generative LLMs with fine-tuned sentence transformers in hybrid systems has further improved intent detection accuracy and efficiency through adaptive in-context learning and chain-of-thought prompting [24].

LLMs excel in generating context-aware and theme-aware keywords, as demonstrated by LLM-TAKE, which significantly outperforms traditional keyword extraction models [2]. Joint models that exploit the relationship between Slot Filling (SF) and Intent Classification (IC) show superior performance compared to independent models, achieving near 'solved' status on benchmark datasets like ATIS and SNIPS [28].

However, experimental findings indicate that LLMs struggle to converge toward market equilibrium, contrasting with human trader behavior, revealing limitations in complex economic environments [25].

The capabilities of LLMs in NLP tasks are rapidly evolving, driven by advancements in model architecture, innovative training methodologies, and targeted application strategies. Recent studies emphasize LLMs' effectiveness in keyword extraction, e-commerce applications, and content generation, showcasing their ability to comprehend complex linguistic patterns and generate contextually relevant responses. Ongoing research addresses challenges related to model biases and interpretability while exploring new training techniques to enhance their utility across various domains [5, 12, 47, 2, 1]. These developments promise further enhancements in language processing and understanding, underscoring the transformative potential of LLMs in NLP.

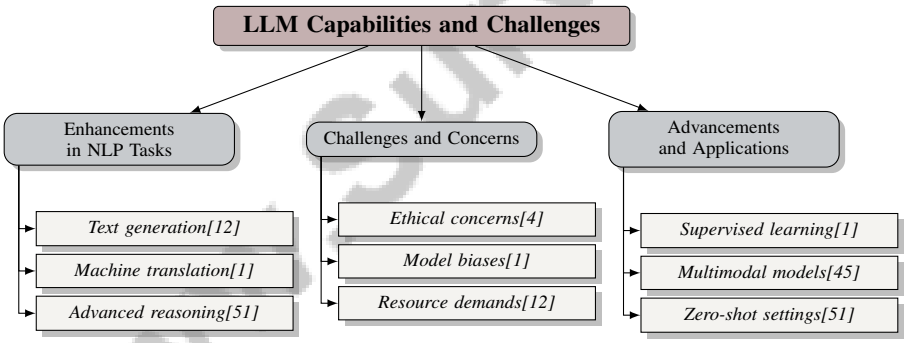


Figure 3: This figure illustrates the capabilities and challenges of Large Language Models (LLMs) in NLP tasks, highlighting enhancements in text generation, machine translation, and reasoning, alongside ethical concerns and resource demands. It also showcases advancements in supervised learning and multimodal models.

3.3 Advantages of LLMs

LLMs offer significant advantages in NLP, primarily through their ability to generate human-like text with high precision and adaptability. A key strength is their capacity to enhance the comprehensibility of generated utterances and improve performance in fulfilling user requests. Hierarchical frameworks like HDNO exemplify this by outperforming existing methods through clearer communication [26].

Integrating LLMs into dialogue systems, particularly in Dialogue State Tracking (DST) tasks, showcases substantial benefits. The LUAS approach reduces data collection costs and time while enhancing task performance, illustrating LLMs' efficiency in streamlining dialogue management processes [27]. This efficiency is essential for developing systems that adapt to diverse conversational contexts and provide accurate, contextually relevant responses.

LLMs automate complex tasks across various fields, including healthcare and e-commerce, by generating contextually relevant content and extracting meaningful keywords. This automation

alleviates the workload for human operators and enhances system reliability through more accurate and consistent outputs. Additionally, advanced evaluation mechanisms for LLM applications ensure that generated content meets critical standards, fostering greater trust in automated solutions [21, 2, 5]. Their adaptability to different scenarios and ability to capture semantic nuances enhance effectiveness in generating activities and supporting personalized user interactions, crucial for fostering user confidence in applications such as writing, translation, and information dissemination.

In healthcare, LLMs significantly improve diagnostic accuracy and patient engagement by integrating domain-specific knowledge into personalized interactions. This capability highlights their transformative potential in revolutionizing healthcare through enhanced clinical decision support, patient engagement, and diagnostic accuracy, ultimately improving care quality and patient outcomes [54, 6, 19, 55, 56]. In conversational AI, LLMs facilitate dynamic multi-agent conversations, improving system performance while reducing code complexity, thus facilitating easier maintenance and scalability.

LLMs continue to push the boundaries of NLP, offering substantial benefits across various tasks and applications. The ongoing development of transformer language models (TLMs) promises to enhance language comprehension and high-quality text generation while enabling seamless integration with specialized domain knowledge. This advancement has the potential to revolutionize human-machine interaction, enhancing multilingual text analytics and opening new research avenues in areas like language user interfaces and collaborative text analysis. As these models evolve, they aim to address current limitations in text mining techniques and contribute to a broader understanding of AI's transformative role in business and society [22, 38, 57, 5].

3.4 Limitations and Challenges

Despite the advanced capabilities of LLMs in NLP, they face significant limitations and challenges hindering widespread application. A primary concern is the high computational cost associated with training and deploying these models, which can be prohibitive for organizations with limited resources. This challenge is exacerbated by the extensive data and computational power required, often restricting access to well-funded entities [14].

Integrating LLMs with multimodal interactions remains underexplored, complicating their deployment in real-world applications [10]. The risk of generating harmful or biased content persists, necessitating ongoing research to mitigate these issues [14].

Evaluation of LLMs is fraught with difficulties, as many studies lack clarity in operationalizing constructs, complicating comparisons across evaluations [58]. Benchmarks often rely on specific LLMs and limited context windows, affecting performance on longer texts or those generated by different models [5]. This limitation underscores the need for comprehensive evaluation metrics that accurately capture model performance across diverse scenarios.

LLMs also struggle with effectively integrating knowledge graphs (KGs), leading to increased complexity and computational costs. This challenge highlights the need for innovative approaches to enhance the models' ability to incorporate external knowledge seamlessly [9]. Moreover, LLMs do not replicate the adaptive learning and emotional factors influencing human behavior, particularly in competitive market environments, leading to erratic outcomes [25].

The generation of inaccurate content, such as keywords, remains a risk due to limitations in LLMs' contextual understanding, impacting effectiveness in tasks requiring precise information retrieval [2]. Ethical concerns and data biases continue to pose significant challenges, affecting the reliability and fairness of LLM-generated outputs [7].

Addressing these limitations requires a multifaceted approach, including novel training methodologies, improved dataset diversity, and more flexible architectures. These efforts are crucial for enhancing the adaptability, efficiency, and robustness of LLMs in complex real-world applications, leveraging advanced techniques such as theme-aware keyword extraction, iterative data enhancement, and retrieval-augmented generation to mitigate common issues like hallucinations, ultimately leading to more effective and scalable AI solutions [59, 2, 5, 60].

As illustrated in Figure 5, the primary limitations and challenges faced by LLMs can be categorized into high computational costs, content and evaluation issues, and knowledge integration challenges. Each category highlights specific concerns such as training resources, bias, evaluation clarity, and the

integration of knowledge graphs, which significantly impact the effectiveness and scalability of LLMs in real-world applications. The integration of embodied and social grounding presents promising avenues for advancing LLM capabilities in meaningful environmental interactions.

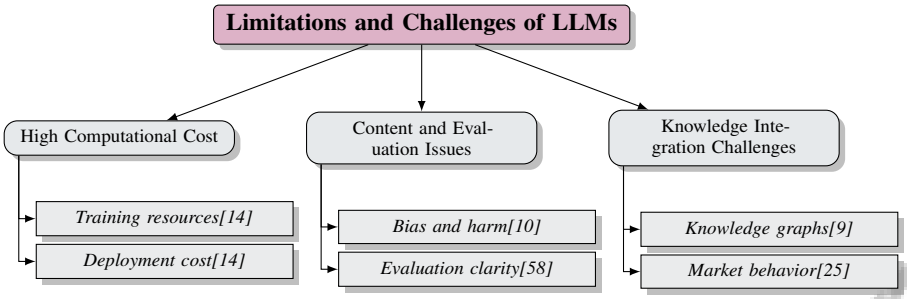


Figure 4: This figure illustrates the primary limitations and challenges faced by Large Language Models (LLMs), categorized into high computational costs, content and evaluation issues, and knowledge integration challenges. Each category highlights specific concerns such as training resources, bias, evaluation clarity, and the integration of knowledge graphs, which impact the effectiveness and scalability of LLMs in real-world applications.

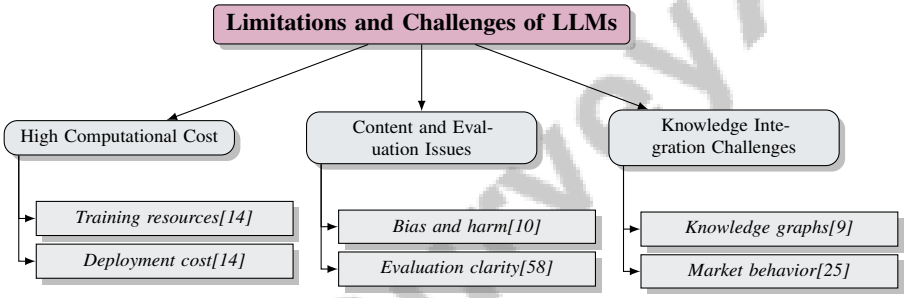


Figure 5: This figure illustrates the primary limitations and challenges faced by Large Language Models (LLMs), categorized into high computational costs, content and evaluation issues, and knowledge integration challenges. Each category highlights specific concerns such as training resources, bias, evaluation clarity, and the integration of knowledge graphs, which impact the effectiveness and scalability of LLMs in real-world applications.

Feature	GPT	BERT	Prompt-based Learning
Training Paradigm	Pre-train, Fine-tune	Pre-train, Fine-tune	Reformulation
Application Domain	Text Generation	Language Understanding	Task Adaptation
Challenges	Ethical Concerns	Resource Demands	Retraining Needs

Table 3: Comparison of Large Language Models (LLMs) in terms of their training paradigms, application domains, and associated challenges. The table highlights the distinct approaches and focuses of GPT, BERT, and prompt-based learning within the field of Natural Language Processing (NLP). This comparative analysis provides insights into the strengths and limitations of each method, aiding in the understanding of their roles and implications in NLP tasks.

4 Task-Oriented Dialogue Systems

4.1 Leveraging LLMs in Task-Oriented Dialogue Systems

The integration of Large Language Models (LLMs) into Task-Oriented Dialogue Systems (TDSs) has revolutionized the management of complex user interactions by enhancing dialogue management with context-aware response generation and sophisticated memory structures. UBAR, for instance, utilizes LLMs to condition responses on all previously generated content, ensuring coherence and contextual awareness [46]. Hybrid approaches combining imitation and reinforcement learning address dialogue

state distribution mismatches, enhancing adaptability and robustness through user feedback [61]. LLMs streamline user interactions and improve information retrieval efficiency, crucial for increasing user satisfaction and task completion rates [24].

Frameworks like Dialog2API leverage LLMs to generate executable programs based on user goals, enabling TDSs to handle more complex interactions [11]. Hierarchical reinforcement learning optimizes dialogue policy and natural language generation asynchronously, enhancing comprehensibility and task performance [26]. The Knowledge Embedded (KE) model integrates knowledge bases directly into model parameters, facilitating efficient dialogue management without external knowledge reliance [30].

LLMs also simulate user-agent interactions, generating annotated dialogue data for Dialogue State Tracking (DST) with models like GPT-4, significantly enhancing system efficiency and accuracy [27]. Neural models excel in capturing complex patterns in Slot Filling (SF) and Intent Classification (IC), achieving state-of-the-art performance on benchmark datasets [28]. Models such as HUS enhance accuracy and contextual appropriateness by encoding user goals and dialogue history to simulate user responses effectively [29].

Despite these advancements, challenges like potential inaccuracies and ethical concerns regarding authorship and intellectual property remain. Thorough assessments of LLM-generated content are necessary to ensure reliability and ethical use. The evolution of LLMs in TDSs reshapes conversational AI, providing sophisticated, context-aware solutions that enhance user satisfaction and task completion rates. Innovations such as adaptive in-context learning and chain-of-thought prompting for intent detection address out-of-scope detection and latency issues, while comprehensive evaluation mechanisms combining human and LLM-based assessments yield insights into improvement areas, ensuring intelligent systems effectively meet user needs across domains like medicine and psychology [21, 23].

4.2 Integration of External Knowledge

Integrating external knowledge into Task-Oriented Dialogue Systems (TDSs) using Large Language Models (LLMs) is crucial for handling complex and dynamic interactions. Memory-augmented dialogue management, as seen in the MAD model, uses a slot-value memory alongside an external memory to capture and apply historical dialogue semantics, significantly enhancing the extraction and application of semantic information for maintaining contextually relevant interactions over extended conversations [62].

As illustrated in Figure 6, the integration of external knowledge in TDSs through LLMs encompasses three main approaches: Memory-Augmented Dialogue, Knowledge Graphs Integration, and Retrieval-Augmented Generation. Each of these approaches enhances the dialogue systems' ability to maintain contextually relevant and factually accurate interactions. LLMs facilitate external knowledge integration by leveraging pre-trained knowledge bases enriched with domain-specific information, essential for tasks requiring current and context-specific knowledge, such as in healthcare or financial services. Incorporating knowledge graphs (KGs) into LLMs enhances their ability to generate accurate and contextually relevant responses, addressing challenges like hallucination and inconsistencies in dialogue outputs. Structured, interconnected factual knowledge from KGs fills gaps in LLM understanding, increasing reliability and applicability across various domains. Research focuses on effective methods for integrating KGs to optimize performance in generating factually grounded content [63, 64, 9].

Retrieval-augmented generation (RAG) techniques enable LLMs to dynamically access external databases and knowledge sources during dialogue, ensuring responses are contextually relevant and factually accurate. These techniques utilize a sophisticated neural architecture that encodes dialogue context and business constraints while employing a two-stage learning strategy to enhance performance [41, 22, 65, 5]. This integration enhances TDS robustness and adaptability, allowing for a broader range of user queries and comprehensive support.

The continued development of LLMs and their integration with external knowledge sources is transforming the dialogue systems landscape, offering sophisticated, context-aware solutions that improve user satisfaction and task completion rates. This integration underscores the importance of combining internal and external knowledge sources to enhance dialogue systems' effectiveness and adaptability, as evidenced by recent advancements employing innovative techniques such as

textual interfaces and domain-specific knowledge injection. These approaches improve the alignment of agent responses with external information and foster more natural interactions and higher task success rates in real-world applications [41, 22, 66, 67].

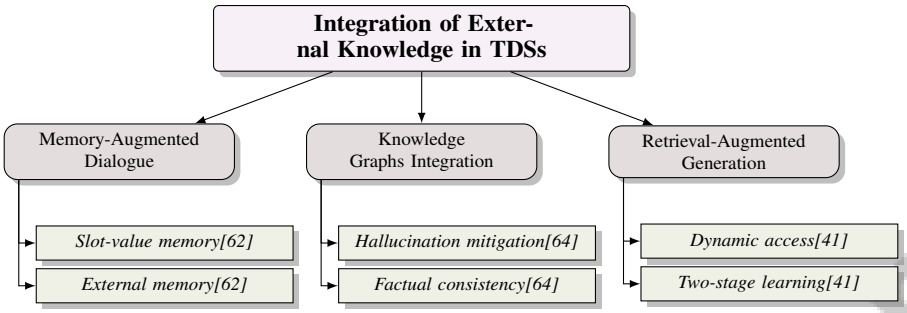


Figure 6: This figure illustrates the integration of external knowledge in Task-Oriented Dialogue Systems (TDSs) using Large Language Models (LLMs), highlighting three main approaches: Memory-Augmented Dialogue, Knowledge Graphs Integration, and Retrieval-Augmented Generation. Each approach enhances the dialogue systems' ability to maintain contextually relevant and factually accurate interactions.

5 Applications of LLMs in Dialogue Systems

The incorporation of Large Language Models (LLMs) into dialogue systems has substantially advanced their capabilities, particularly in natural language understanding and generation. This section delves into the varied applications of LLMs, highlighting their transformative impact across different sectors through innovative implementations and case studies that demonstrate their role in enhancing user interactions and system efficacy.

5.1 Innovative Applications and Case Studies

LLMs have catalyzed significant progress in dialogue systems, as evidenced by diverse applications across numerous domains. In healthcare, LLMs have improved diagnostic processes, notably in ophthalmology, where models like GPT-4 enhance AI application reliability and efficiency [68]. Speech LLMs have also advanced speech recognition accuracy and contextual comprehension, showcasing their proficiency in managing multimodal inputs.

In business process management, LLMs automate content generation, exemplified by Tweet Writer, which creates tweets for real-time alerts, boosting user engagement and operational efficiency [69]. The residual adapter model further demonstrates LLM adaptability by managing dialogue responses across various styles and topics while ensuring fluency and coherence [70].

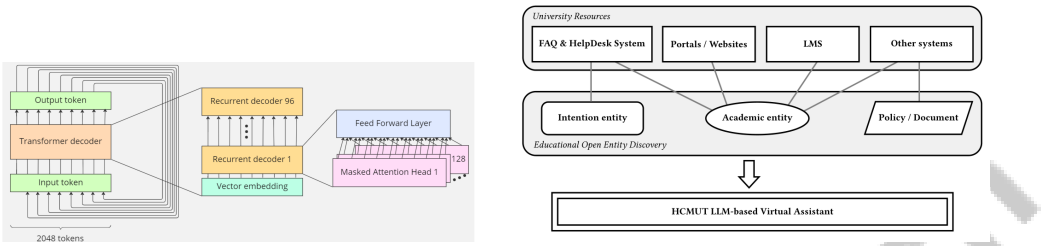
The SAMIA framework enhances dialogue quality and user experience by extending deep reinforcement learning capabilities in task-oriented dialogue systems [71]. The GoEx framework improves user trust in LLM applications by providing a safety net through undo functionality and damage confinement [72].

In task-oriented dialogue systems, the MTTOD model has set new benchmarks, scoring 108.3 on MultiWOZ 2.0 and 107.5 on MultiWOZ 2.1, surpassing other models in success rates and BLEU scores [34]. This model exemplifies LLM potential to enhance dialogue success rates and overall performance, even in unsupervised settings [73].

In e-commerce, LLMs improve shopping experiences through applications like LLM-TAKE, which enhances keyword extraction accuracy and diversity [2]. This application highlights LLM effectiveness in facilitating online shopping.

Emerging benchmarks now incorporate user follow-up utterances into dialogue system evaluations, contrasting with traditional methods that rely solely on initial queries, emphasizing the importance of continuous interaction in assessing dialogue system performance [74].

The case studies presented reveal the significant applications of LLMs in dialogue systems, showcasing their transformative potential across sectors such as healthcare and e-commerce. These studies illustrate how LLMs foster rapid development of specialized chatbots and enhance keyword extraction through advanced contextual understanding, while emphasizing the need for rigorous evaluation methods to ensure the effectiveness and reliability of these technologies in improving user experiences and driving innovation [21, 2].



(a) Transformer Recurrent Decoder Architecture[40]

(b) Educational Open Entity Discovery System[75]

Figure 7: Examples of Innovative Applications and Case Studies

As shown in Figure 7, LLMs have emerged as pivotal tools in dialogue systems, driving innovation and enhancing user interaction across applications. The first example, the "Transformer Recurrent Decoder Architecture," combines transformer and recurrent decoder layers to optimize input and output token processing, illustrating how LLMs can enhance dialogue generation and understanding. The second example, the "Educational Open Entity Discovery System," demonstrates LLM application in education, where a virtual assistant aids users in navigating university resources. By integrating systems such as FAQs and Learning Management Systems, this application underscores LLMs' transformative potential in creating cohesive educational environments. Together, these examples highlight LLMs' diverse and impactful contributions to advancing dialogue systems.

5.2 Domain-Specific Enhancements

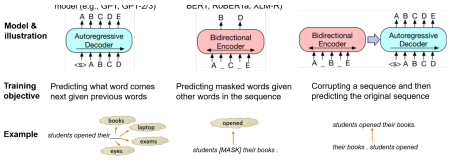
LLMs have been increasingly customized to meet the specific needs of various domains, such as healthcare and e-commerce, leveraging domain-specific data to enhance performance. In healthcare, LLMs improve diagnostic accuracy and patient interaction by integrating medical knowledge bases, facilitating applications like automated diagnosis [24].

In e-commerce, LLMs enhance the shopping experience by providing personalized recommendations and improving customer service interactions. Models like LLM-TAKE generate contextually relevant keywords, significantly improving keyword extraction accuracy and diversity [2]. These advancements illustrate LLMs' potential to transform e-commerce platforms by automating content generation and optimizing customer engagement.

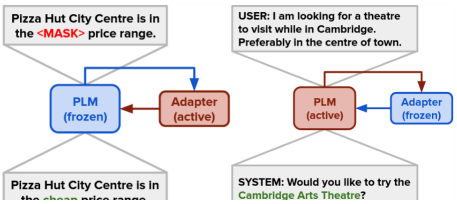
The creation of datasets like ALLWOZ, comprising customer service dialogs across various domains, exemplifies enhancements for multilingual task-oriented dialogue systems. This dataset highlights the importance of tailoring LLMs to handle linguistic and contextual nuances, improving effectiveness in real-world applications [76].

Domain-specific enhancements of LLMs showcase their versatility across sectors, particularly in applications such as keyword extraction and optimization. These advancements enable tailored language processing solutions that meet unique domain requirements, exemplified by the Theme-Aware Keyword Extraction framework, which improves keyword generation through contextual metadata [47, 2, 77].

As shown in Figure 8, domain-specific enhancements are crucial for tailoring LLMs to meet the nuanced needs of particular domains. The first figure illustrates three sequence modeling tasks—autoregressive decoding, bidirectional encoding, and autoregressive decoding with masking—designed to optimize information flow between the model and input sequence, demonstrating LLM versatility in handling various sequence data. The second figure presents a user's query for a theatre in Cambridge, emphasizing the importance of integrating domain-specific knowledge into



(a) Model illustration of training objective and example for different types of sequence modeling tasks[38]



(b) User's query for a theatre in Cambridge, with a preference for the centre of town[66]

Figure 8: Examples of Domain-Specific Enhancements

LLMs to process and respond accurately, thereby enhancing user experience in real-world dialogue systems.

5.3 Multilingual and Cross-Domain Applications

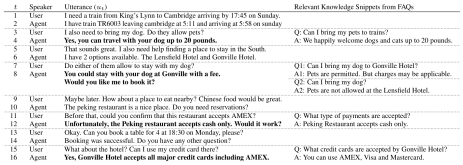
The application of LLMs in multilingual and cross-domain dialogue systems represents a transformative approach to enhancing conversational AI capabilities. LLMs facilitate seamless interactions in multilingual environments by understanding and generating language across diverse contexts without extensive language-specific training data [24].

In multilingual dialogue systems, LLMs maintain coherence and contextual relevance across languages, essential for global market applications where users expect consistent interactions. The integration of LLMs not only enhances user satisfaction but broadens access to digital services for non-English speakers, democratizing information access [9].

Cross-domain applications further illustrate LLM versatility, enabling seamless transitions between domains such as healthcare, finance, and customer service. Their adaptability is supported by extensive pre-training on diverse datasets, equipping them with a broad understanding of various contexts and terminologies [2]. This adaptability is particularly beneficial for developing task-oriented dialogue systems that require integration with domain-specific knowledge bases [9].

The development of multilingual and cross-domain dialogue systems is enhanced by knowledge graphs (KGs) and retrieval-augmented generation (RAG) techniques, allowing LLMs to access and integrate external information dynamically. This ensures that responses are linguistically accurate and factually grounded, improving system reliability and effectiveness [9].

The integration of LLMs in multilingual and cross-domain dialogue systems marks a significant advancement in conversational AI. These systems not only enhance inclusivity by accommodating diverse linguistic backgrounds but also increase versatility across industries such as education, healthcare, and customer service. This evolution is driven by the vast amounts of data generated globally, which fuels LLM training, enabling the generation of contextually relevant and nuanced responses. Consequently, LLM-based chatbots are setting new benchmarks in AI technology, paving the way for more effective conversational agents that cater to a broad spectrum of users [21, 7]. Ongoing development and refinement of these models promise to further enhance their capabilities, leading to more sophisticated and user-centric dialogue systems.



t	Speaker	Utterance (u_t)	Relevant Knowledge Snippets from FAQs
1	User	I need a train from King's Lynn to Cambridge arriving by 17:45 on Sunday.	
2	Agent	I have train TR6003 leaving cambridge at 5:11 and arriving at 5:58 on sunday	
3	User	I also need to bring my dog. Do they allow pets?	Q: Can I bring my pets to trains?
4	Agent	**Yes, you can travel with your dog up to 20 pounds.**	A: We happily welcome dogs and cats up to 20 pounds.
5	User	That sounds great. I also need help finding a place to stay in the South.	
6	Agent	I have 2 options available. The Lensfield Hotel and Gonville Hotel.	
7	User	Do either of them allow to stay with my dog?	Q1: Can I bring my dog to Gonville Hotel?
8	Agent	**You could stay with your dog at Gonville with a fee. Would you like me to book it?**	A1: Pets are permitted. But charges may be applicable. Q2: Can I bring my dog? A2: Pets are not allowed at the Lensfield Hotel.
9	User	Maybe later. How about a place to eat nearby? Chinese food would be great.	
10	Agent	The peking restaurant is a nice place. Do you need reservations?	
11	User	Before that, could you confirm that this restaurant accepts AMEX?	Q: What type of payments are accepted?
12	Agent	**Unfortunately, the Peking restaurant accepts cash only. Would it work?**	A: Peking Restaurant accepts cash only.
13	User	Okay. Can you book a table for 4 at 18:30 on Monday, please?	
14	Agent	Booking was successful. Do you have any other question?	
15	User	What about the hotel? Can I use my credit card there?	Q: What credit cards are accepted by Gonville Hotel?
16	Agent	**Yes, Gonville Hotel accepts all major credit cards including AMEX.**	A: You can use AMEX, Visa and Mastercard.

(a) The image shows a table with a list of utterances and relevant knowledge snippets from FAQs.[78]

(b) Translation and Human Editing Process[76]

Figure 9: Examples of Multilingual and Cross-Domain Applications

AI-generated, for reference only.

As shown in Figure 9, LLMs demonstrate remarkable potential in multilingual and cross-domain applications. The first image illustrates a table organizing conversational data, featuring utterances alongside corresponding knowledge snippets from FAQs, showcasing how LLMs integrate domain-specific knowledge to enhance response relevance and accuracy. The second image outlines a translation and human editing process, emphasizing LLMs' role in bridging language barriers, demonstrating the importance of combining machine-driven processes with human expertise to ensure high-quality translations. Together, these examples underscore LLMs' versatility in advancing dialogue systems across linguistic and contextual boundaries.

6 Conversational AI and Dialogue Management

6.1 Enhancing Conversational Flow and Context Management

Large Language Models (LLMs) have significantly advanced conversational flow and context management, offering sophisticated mechanisms for managing complex dialogues across domains. By using structured prompts, LLMs enhance interaction quality, crucial in specialized fields like clinical decision-making where context-aware communication is essential [55]. This capability ensures coherence and relevance in extended dialogues, boosting user satisfaction and task completion rates.

LLMs' adaptability is exemplified by tools like NeMo Guardrails, which allow real-time adjustment of conversational boundaries, essential for managing dynamic interactions in complex conversations [79]. Interdisciplinary collaboration among ethicists, technologists, and policymakers is vital to align LLM advancements with ethical standards, addressing the implications of their deployment and promoting responsible development [80].

Enhancements in conversational flow and context management underscore LLMs' transformative potential in AI, facilitating coherent, contextually aware interactions. This significantly enriches user experiences, particularly in medicine and psychology, where LLM-based chatbots are increasingly used. However, rigorous evaluation of these systems is crucial to address discrepancies between automated and human assessments, ensuring reliability and effectiveness in generating knowledge and responding to user prompts [21, 7].

6.2 Advanced Dialogue Management

The incorporation of LLMs into dialogue management systems has enhanced the field, enabling sophisticated conversational strategies. These models support systems that dynamically adapt to user inputs, managing complex interactions with improved precision and contextual understanding. A key innovation is hierarchical reinforcement learning, which optimizes dialogue policies across multiple abstraction levels, enhancing dialogue flow management through learned strategies for both high-level acts and low-level generation tasks [26].

LLMs also facilitate the integration of external knowledge sources, such as knowledge graphs and retrieval-augmented generation techniques, enriching dialogue management by dynamically incorporating relevant information for contextually appropriate and factually accurate responses [9, 24]. Advanced management techniques benefit from meta-learning strategies, enabling LLMs to quickly adapt to new tasks and domains with minimal data, as demonstrated by approaches like Meta-NLG, which simulate multiple tasks during training [43].

6.3 Structural Approaches to Dialogue Management

Structural approaches to dialogue management in conversational AI have evolved with LLMs, enhancing the ability to manage complex interactions efficiently. A modular architecture segments dialogue management into components like Natural Language Understanding, Dialogue State Tracking, Dialogue Policy, and Natural Language Generation, allowing independent optimization and tailored enhancements across scenarios [81, 22, 82, 77].

Hierarchical frameworks, such as hierarchical reinforcement learning, organize dialogue management tasks at multiple levels, optimizing strategies for both high-level acts and low-level generation tasks, improving system efficiency [26]. The integration of external knowledge sources like knowledge graphs and retrieval-augmented generation techniques ensures responses are contextually and factually accurate, enhancing comprehensive response capabilities across domains [9].

Meta-learning strategies further benefit structural approaches, allowing systems to adapt swiftly to new tasks and domains with minimal data. Methods like Meta-NLG simulate multiple meta tasks during training, facilitating rapid adaptation to novel contexts, crucial for maintaining high performance across diverse applications [43].

7 Challenges and Future Directions

7.1 Ethics and Vulnerabilities in Conversational AI

Deploying Large Language Models (LLMs) in conversational AI systems necessitates a thorough examination of ethical considerations and inherent vulnerabilities. A primary ethical concern is the presence of biases in training data, which can skew outputs and adversely affect decision-making in sensitive areas like healthcare, underscoring the need for responsible AI deployment to maintain trust in LLM systems [24]. Furthermore, inaccuracies in LLM-generated content, such as belief states and system acts, can detrimentally impact dialogue quality [46].

The limitations of LLMs in contextual understanding pose risks, particularly in tasks that require precise information retrieval [2]. This issue is exacerbated by the models' varying capabilities in handling multilingual data and complex user requests, which can introduce biases in AI outputs [4]. Additionally, discrepancies in quality between LLM-generated and manually annotated data raise ethical concerns, especially in dialogue systems where accuracy is vital [27].

While hierarchical frameworks for dialogue management enhance performance, they still face challenges in specific contexts or with complex requests, necessitating ongoing evaluation and refinement [26]. The ethical implications of these limitations highlight the need for robust verification mechanisms and fallback strategies to mitigate risks associated with LLM-generated content.

Future research should focus on enhancing the interpretability of LLM behaviors, developing systematic evaluation frameworks, and exploring methodologies to reduce biases and inaccuracies. Aligning LLM development with ethical standards and societal needs is crucial for addressing challenges such as hallucination, accountability, and bias, while promoting transparency in applications. Implementing tailored ethical frameworks and dynamic auditing systems will mitigate risks and foster interdisciplinary collaboration among stakeholders in academia, industry, and civil society, ultimately maximizing the utility of LLMs across diverse applications [80, 83, 39].

7.2 Computational and Resource Constraints

The deployment of Large Language Models (LLMs) in dialogue systems faces significant computational and resource constraints, hindering scalability and efficiency. Traditional training methods incur high computational costs, necessitating extensive data collection and fine-tuning, which often restricts deployment to well-resourced organizations [84]. The ANNM model exemplifies the substantial resources and training data required, complicating the deployment of advanced models in real-world applications [85].

Innovative approaches, such as prompt engineering, have been proposed to alleviate the need for extensive data collection and fine-tuning [86]. The ALLIES method emphasizes the high costs associated with multiple API calls, highlighting the necessity for more efficient computational strategies [87]. Additionally, the reliance of the DAUS model on LLMs, which can produce hallucinations and lack open-source availability, affects replicability and generalizability [88].

Automated solutions, such as the AutoFlow framework, offer a means to address computational costs linked to manually designed workflows, proposing an automated approach to mitigate these constraints [89]. Furthermore, employing retrieval-augmented generation (RAG) across diverse languages and contexts can optimize retrieval mechanisms and enhance data quality, contributing to system robustness [60].

Future research should aim to optimize the computational efficiency of adapters and explore additional attributes for control in dialogue systems [70]. Efforts to reduce resource consumption, as suggested by recent studies, positively impact energy savings and service deployment [90]. Additionally, addressing challenges related to generating accurate responses for infrequent entities and managing real-world data noise will necessitate ongoing refinement of LLM methodologies.

Tackling the computational and resource constraints of LLMs requires a multifaceted approach, including advancements in storage and processing technologies, innovative training methodologies, and the development of more efficient architectures. Such efforts are essential for enhancing the scalability and accessibility of LLMs across various applications and industries, ensuring the efficient use of limited device resources [91].

7.3 Language and Data Biases

Large Language Models (LLMs) exhibit remarkable capabilities in processing and generating human language, yet they are susceptible to biases inherent in their training data. These biases, reflecting societal prejudices, can manifest in forms such as racial, gender, and cultural biases, leading to skewed outputs and the reinforcement of stereotypes. Addressing language and data biases in LLMs is further complicated by variability in model performance across tasks, necessitating robust evaluation frameworks to ensure fairness and accuracy in applications [92].

Effective evaluation of LLMs for language and data biases requires comprehensive benchmarks that assess performance across diverse linguistic contexts and domains. Such benchmarks are vital for identifying and mitigating biases stemming from training data, which often lacks representation from underrepresented groups and languages. The LogEval benchmark suite exemplifies efforts to provide robust evaluation tools for log analysis, emphasizing the need to evaluate LLMs in a manner that accounts for potential biases [92].

Addressing language and data biases involves improving the diversity and quality of training datasets and developing methodologies to identify and correct biased outputs. Techniques such as debiasing algorithms and fairness-aware training protocols are essential for enhancing the ethical deployment of LLMs, ensuring they do not perpetuate harmful biases. Research into the interpretability of LLMs is crucial for uncovering how biases are embedded within these systems. Understanding these processes enables the development of targeted strategies to mitigate biases, thereby enhancing the ethical use of LLMs across various applications, including natural language processing, scientific writing, and content generation. This understanding is vital for fostering transparency and accountability in AI technologies, addressing both ethical concerns and the practical implications of bias in real-world scenarios [93, 5, 17, 39, 2].

7.4 Future Directions and Research Opportunities

The advancement of Large Language Models (LLMs) in dialogue systems presents numerous research opportunities aimed at overcoming existing challenges and enhancing applicability across diverse contexts. A significant area of exploration involves optimizing the efficiency and scalability of LLMs, particularly in task-oriented dialogue systems. Future research should prioritize enhancing the quality of data generated by LLMs, investigating the applicability of LUAS (a universal analysis framework for LLMs) to a broader range of tasks beyond Dialogue State Tracking (DST), and strengthening user simulation models to accommodate diverse user personalities. This approach will improve AI system robustness and reliability while addressing emerging trustworthiness concerns [22, 94, 40, 83].

In multilingual and cross-domain applications, developing portable models that effectively operate across different domains and languages is crucial. This requires leveraging unlabeled data and exploring generative models to improve evaluation metrics for complex utterances, ensuring dialogue systems adapt to a wide range of linguistic and contextual nuances [28]. Additionally, integrating adaptive learning and behavioral economics into LLMs could enhance their effectiveness in economic modeling and simulations, providing more accurate and contextually aware responses [25].

The exploration of Knowledge Graph-Enhanced LLMs (KGLLMs) represents another promising research avenue. Future studies should focus on improving KGLLM efficiency, exploring multi-modal and temporal knowledge graphs, and enhancing interpretability. Developing domain-specific KGLLMs will address current model limitations and facilitate more accurate interactions [9]. Furthermore, refining the LLM-TAKE framework to enhance robustness against hallucinations is essential for improving keyword extraction processes [2].

Additionally, future research should investigate broader applications of AI-assisted approaches in data management, focusing on methodologies to improve accuracy in multilingual data processing [4]. Enhancing the scalability and accessibility of LLMs across various applications and industries

will ensure efficient use of limited resources, leading to more sophisticated, efficient, and user-centric AI systems.

By proactively addressing identified future research directions, scholars can significantly enhance the functionality and effectiveness of LLMs. This enhancement is crucial for the ongoing advancement of LLM technology and for meeting the increasing demands of diverse applications, particularly in dialogue systems and other fields. Such efforts will involve careful evaluation of ethical considerations, bias mitigation, and the establishment of guidelines promoting responsible use, ensuring LLMs contribute positively to research and education while minimizing potential deployment risks [22, 39].

8 Conclusion

The exploration of Large Language Models (LLMs) within this survey underscores their profound impact on Natural Language Processing (NLP) and Task-Oriented Dialogue Systems (TDSs), marking a significant leap in artificial intelligence research and practical applications. The incorporation of models like ChatGPT has notably improved system resilience and user interaction, while also reducing the development burden. Despite these advancements, there remain substantial challenges, particularly concerning the inherent limitations and biases of LLMs, necessitating ongoing innovation.

Developing a comprehensive framework for evaluating dialogue systems is imperative, integrating various constructs and metrics to thoroughly measure user experience and system performance. Such a framework is crucial for ensuring the dependability and effectiveness of dialogue systems, as demonstrated by systems like Dialog-To-Actions (DTA) that address challenges linked to end-to-end generation techniques. Additionally, ethical considerations in LLM deployment are paramount, requiring robust datasets and ethical research methodologies to mitigate biases and ensure equitable and responsible AI applications.

As LLMs continue to evolve, their ability to automate intricate tasks and enhance efficiency, particularly in sectors like telecommunications, becomes increasingly apparent. This highlights the importance of developing tailored strategies to fully harness their capabilities. The survey further highlights the potential of integrating LLMs with Knowledge Graphs (KGs) to enhance their functionality, improving both interpretability and factual accuracy, which are critical for applications in high-stakes scenarios. Moreover, advancements in tool learning have shown to significantly augment LLM capabilities in handling complex queries, thus enhancing accuracy, efficiency, and user trust.

References

[1] Muhammad Usman Hadi, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, et al. A survey on large language models: Applications, challenges, limitations, and practical usage. *Authorea Preprints*, 3, 2023.

[2] Reza Yousefi Maragheh, Chenhao Fang, Charan Chand Irugu, Parth Parikh, Jason Cho, Jianpeng Xu, Saranyan Sukumar, Malay Patel, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. Llm-take: Theme aware keyword extraction using large language models, 2023.

[3] Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*, 2022.

[4] Juhani Merilehto. From pdfs to structured data: Utilizing llm analysis in sports database management, 2024.

[5] Irina Tolstykh, Aleksandra Tsybina, Sergey Yakubson, Aleksandr Gordeev, Vladimir Dokholyan, and Maksim Kuprashevich. Gigacheck: Detecting llm-generated content, 2024.

[6] Yining Huang, Keke Tang, Meilian Chen, and Boyuan Wang. A comprehensive survey on evaluating large language model applications in the medical industry, 2024.

[7] Sumit Kumar Dam, Choong Seon Hong, Yu Qiao, and Chaoning Zhang. A complete survey on llm-based ai chatbots. *arXiv preprint arXiv:2406.16937*, 2024.

[8] Tom Bocklisch, Thomas Werkmeister, Daksh Varshneya, and Alan Nichol. Task-oriented dialogue with in-context learning. *arXiv preprint arXiv:2402.12234*, 2024.

[9] Linyao Yang, Hongyang Chen, Zhao Li, Xiao Ding, and Xindong Wu. Give us the facts: Enhancing large language models with knowledge graphs for fact-aware language modeling, 2024.

[10] Yuyou Gan, Yong Yang, Zhe Ma, Ping He, Rui Zeng, Yiming Wang, Qingming Li, Chunyi Zhou, Songze Li, Ting Wang, Yunjun Gao, Yingcai Wu, and Shouling Ji. Navigating the risks: A survey of security, privacy, and ethics threats in llm-based agents, 2024.

[11] Seokhwan Kim, Mihail Eric, Behnam Hedayatnia, Karthik Gopalakrishnan, Yang Liu, Chao-Wei Huang, and Dilek Hakkani-Tur. Beyond domain apis: Task-oriented conversational modeling with unstructured knowledge access track in dstc9, 2021.

[12] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*, 2023.

[13] Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly, 2024.

[14] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

[15] Yuanfeng Song, Yuanqin He, Xuefang Zhao, Hanlin Gu, Di Jiang, Haijun Yang, Lixin Fan, and Qiang Yang. A communication theory perspective on prompting engineering methods for large language models, 2023.

[16] John Mendonça, Alon Lavie, and Isabel Trancoso. On the benchmarking of llms for open-domain dialogue evaluation, 2024.

[17] Morgan Fouesneau, Ivelina G. Momcheva, Urmila Chadayammuri, Mariia Demianenko, Antoine Dumont, Raphael E. Hviding, K. Angelique Kahle, Nadiia Pulatova, Bhavesh Rajpoot, Marten B. Scheuck, Rhys Seeburger, Dmitry Semenov, and Jaime I. Villaseñor. What is the role of large language models in the evolution of astronomy research?, 2024.

[18] Changle Qu, Sunhao Dai, Xiaochi Wei, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, Jun Xu, and Ji-Rong Wen. Tool learning with large language models: A survey, 2024.

[19] Qian Niu, Keyu Chen, Ming Li, Pohsun Feng, Ziqian Bi, Lawrence KQ Yan, Yichao Zhang, Caitlyn Heqi Yin, Cheng Fei, Junyu Liu, Benji Peng, Tianyang Wang, Yunze Wang, Silin Chen, and Ming Liu. From text to multimodality: Exploring the evolution and impact of large language models in medical practice, 2024.

[20] Hengchao Shang, Zongyao Li, Jiaxin Guo, Shaojun Li, Zhiqiang Rao, Yuanchang Luo, Daimeng Wei, and Hao Yang. An end-to-end speech summarization using large language model, 2024.

[21] Bhashithe Abeysinghe and Ruhan Circi. The challenges of evaluating llm applications: An analysis of automated, human, and llm-based approaches, 2024.

[22] Andreas Jungherr. Using chatgpt and other large language model (llm) applications for academic paper assignments. 2023.

[23] Gaurav Arora, Shreya Jain, and Srujana Merugu. Intent detection in the age of llms, 2024.

[24] Cheonsu Jeong. Generative ai service implementation using llm application architecture: based on rag model and langchain framework. *Journal of Intelligence and Information Systems*, 29(4):129–164, 2023.

[25] Jingru Jia and Zehua Yuan. An experimental study of competitive market behavior through llms, 2024.

[26] Jianhong Wang, Yuan Zhang, Tae-Kyun Kim, and Yunjie Gu. Modelling hierarchical structure between dialogue policy and natural language generator with option framework for task-oriented dialogue system, 2021.

[27] Cheng Niu, Xingguang Wang, Xuxin Cheng, Juntong Song, and Tong Zhang. Enhancing dialogue state tracking models through llm-backed user-agents simulation, 2024.

[28] Samuel Louvan and Bernardo Magnini. Recent neural methods on slot filling and intent classification for task-oriented dialogue systems: A survey, 2020.

[29] User modeling for task oriented.

[30] Andrea Madotto, Samuel Cahyawijaya, Genta Indra Winata, Yan Xu, Zihan Liu, Zhaojiang Lin, and Pascale Fung. Learning knowledge bases with parameters for task-oriented dialogue systems, 2020.

[31] Linmei Hu, Zeyi Liu, Ziwang Zhao, Lei Hou, Liqiang Nie, and Juanzi Li. A survey of knowledge enhanced pre-trained language models, 2023.

[32] Victor Uc-Cetina, Nicolas Navarro-Guerrero, Anabel Martin-Gonzalez, Cornelius Weber, and Stefan Wermter. Survey on reinforcement learning for language processing, 2022.

[33] Evgeniia Razumovskaia, Goran Glavaš, Olga Majewska, Edoardo M. Ponti, Anna Korhonen, and Ivan Vulić. Crossing the conversational chasm: A primer on natural language processing for multilingual task-oriented dialogue systems, 2022.

[34] Improving end-to-end task-orient.

[35] Songbo Hu, Xiaobin Wang, Zhangdie Yuan, Anna Korhonen, and Ivan Vulić. Dialight: Lightweight multilingual development and evaluation of task-oriented dialogue systems with large language models, 2024.

[36] Yuncheng Hua, Xiangyu Xi, Zheng Jiang, Guanwei Zhang, Chaobo Sun, Guanglu Wan, and Wei Ye. Dialog-to-actions: Building task-oriented dialogue system via action-level generation, 2023.

[37] Shivom Aggarwal, Shourya Mehra, and Pritha Mitra. Multi-purpose nlp chatbot : Design, methodology conclusion, 2023.

[38] Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40, 2023.

[39] Ahmed S. BaHammam, Khaled Trabelsi, Seithikurippu R. Pandi-Perumal, and Hiatham Jahrami. Adapting to the impact of ai in scientific writing: Balancing benefits and drawbacks while developing policies and regulations, 2023.

[40] Frank Joublin, Antonello Ceravola, Joerg Deigmoeller, Michael Gienger, Mathias Franzius, and Julian Eggert. A glimpse in chatgpt capabilities and its impact for ai research, 2023.

[41] Lahari Poddar, György Szarvas, Cheng Wang, Jorge Balazs, Pavel Danchenko, and Patrick Ernst. Deploying a retrieval based response model for task oriented dialogues, 2022.

[42] Weizhi Wang, Zhirui Zhang, Junliang Guo, Yinpei Dai, Boxing Chen, and Weihua Luo. Task-oriented dialogue system as natural language generation. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pages 2698–2703, 2022.

[43] Fei Mi, Minlie Huang, Jiyong Zhang, and Boi Faltings. Meta-learning for low-resource natural language generation in task-oriented dialogue systems, 2019.

[44] Zongyue Qin, Chen Luo, Zhengyang Wang, Haoming Jiang, and Yizhou Sun. Relational database augmented large language model, 2024.

[45] Hanyao Huang, Ou Zheng, Dongdong Wang, Jiayi Yin, Zijin Wang, Shengxuan Ding, Heng Yin, Chuan Xu, Renjie Yang, Qian Zheng, and Bing Shi. Chatgpt for shaping the future of dentistry: The potential of multi-modal large language model, 2023.

[46] Yunyi Yang, Yunhao Li, and Xiaojun Quan. Ubar: Towards fully end-to-end task-oriented dialog system with gpt-2. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14230–14238, 2021.

[47] Chester Palen-Michel, Ruixiang Wang, Yipeng Zhang, David Yu, Canran Xu, and Zhe Wu. Investigating llm applications in e-commerce, 2024.

[48] Jingying Zeng, Richard Huang, Waleed Malik, Langxuan Yin, Bojan Babic, Danny Shacham, Xiao Yan, Jaewon Yang, and Qi He. Large language models for social networks: Applications, challenges, and solutions, 2024.

[49] Sara Incao, Carlo Mazzola, Giulia Belgiovine, and Alessandra Sciutti. A roadmap for embodied and social grounding in llms, 2024.

[50] Bosheng Ding, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie Hu, Anh Tuan Luu, and Shafiq Joty. Data augmentation using large language models: Data perspectives, learning paradigms and challenges, 2024.

[51] Language models as few-shot lear.

[52] Ryan Fellows, Hisham Ihshaish, Steve Battle, Ciaran Haines, Peter Mayhew, and J Ignacio Deza. Task-oriented dialogue systems: performance vs. quality-optima, a review. *arXiv preprint arXiv:2112.11176*, 2021.

[53] Taaha Kazi, Ruiliang Lyu, Sizhe Zhou, Dilek Hakkani-Tur, and Gokhan Tur. Large language models as user-agents for evaluating task-oriented-dialogue systems, 2024.

[54] Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerova, Nidhi Rohatgi, Poonam Hosamani, William Collins, Neera Ahuja, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, John Pauly, and Akshay S. Chaudhari. Adapted large language models can outperform medical experts in clinical text summarization, 2024.

[55] Fatemeh Nazary, Yashar Deldjoo, Tommaso Di Noia, and Eugenio di Sciascio. Xai4llm. let machine learning models and llms collaborate for enhanced in-context learning in healthcare, 2024.

[56] Hanguang Xiao, Feizhong Zhou, Xingyue Liu, Tianqi Liu, Zhipeng Li, Xin Liu, and Xiaoxuan Huang. A comprehensive survey of large language models and multimodal large language models in medicine, 2024.

[57] Ross Gruetzemacher and David Paradice. Deep transfer learning beyond: Transformer language models in information systems research, 2021.

[58] Anouck Braggaar, Christine Liebrecht, Emiel van Miltenburg, and Emiel Krahmer. Evaluating task-oriented dialogue systems: A systematic review of measures, constructs and their operationalisations. *arXiv preprint arXiv:2312.13871*, 2023.

[59] Nicholas Lee, Thanakul Wattanawong, Sehoon Kim, Karttikeya Mangalam, Sheng Shen, Gopala Anumanchipalli, Michael W. Mahoney, Kurt Keutzer, and Amir Gholami. Llm2llm: Boosting llms with novel iterative data enhancement, 2024.

[60] Cheonsu Jeong. A study on the implementation of generative ai services using an enterprise data-based llm application architecture. *arXiv preprint arXiv:2309.01105*, 2023.

[61] Xiangkun Hu, Junqi Dai, Hang Yan, Yi Zhang, Qipeng Guo, Xipeng Qiu, and Zheng Zhang. Dialogue meaning representation for task-oriented dialogue systems, 2022.

[62] Zheng Zhang, Minlie Huang, Zhongzhou Zhao, Feng Ji, Haiqing Chen, and Xiaoyan Zhu. Memory-augmented dialogue management for task-oriented dialogue systems, 2018.

[63] Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

[64] Ernests Lavrinovics, Russa Biswas, Johannes Bjerva, and Katja Hose. Knowledge graphs, large language models, and hallucinations: An nlp perspective, 2024.

[65] Xue-Yong Fu, Md Tahmid Rahman Laskar, Cheng Chen, and Shashi Bhushan TN. Are large language models reliable judges? a study on the factuality evaluation capabilities of llms, 2023.

[66] Denis Emelin, Daniele Bonadiman, Sawsan Alqahtani, Yi Zhang, and Saab Mansour. Injecting domain knowledge in language models for task-oriented dialogue systems, 2022.

[67] Qingyang Wu, Deema Alnuhait, Derek Chen, and Zhou Yu. Using textual interface to align external knowledge for end-to-end task-oriented dialogue systems, 2023.

[68] Ting Fang Tan, Kabilan Elangovan, Liyuan Jin, Yao Jie, Li Yong, Joshua Lim, Stanley Poh, Wei Yan Ng, Daniel Lim, Yuhe Ke, Nan Liu, and Daniel Shu Wei Ting. Fine-tuning large language model (llm) artificial intelligence chatbots in ophthalmology and llm-based evaluation using gpt-4, 2024.

[69] Jiahao Wang and Amer Shalaby. Leveraging large language models for enhancing public transit services, 2024.

[70] Andrea Madotto. Taming the beast: Learning to control neural conversational models, 2021.

[71] Weiyan Wang, Yuxiang WU, Yu Zhang, Zhongqi Lu, Kaixiang Mo, and Qiang Yang. Integrating user and agent models: A deep task-oriented dialogue system, 2017.

[72] Shishir G. Patil, Tianjun Zhang, Vivian Fang, Noppapon C., Roy Huang, Aaron Hao, Martin Casado, Joseph E. Gonzalez, Raluca Ada Popa, and Ion Stoica. Goex: Perspectives and designs towards a runtime for autonomous llm applications, 2024.

[73] Brendan King and Jeffrey Flanigan. Unsupervised end-to-end task-oriented dialogue with llms: The power of the noisy channel, 2024.

[74] Clemencia Siro, Mohammad Aliannejadi, and Maarten de Rijke. Rethinking the evaluation of dialogue systems: Effects of user feedback on crowdworkers and llms, 2024.

[75] Tuan Bui, Oanh Tran, Phuong Nguyen, Bao Ho, Long Nguyen, Thang Bui, and Tho Quan. Cross-data knowledge graph construction for llm-enabled educational question-answering system: A case study at hcmut, 2024.

[76] Lei Zuo, Kun Qian, Bowen Yang, and Zhou Yu. Allwoz: Towards multilingual task-oriented dialog systems for all. *arXiv preprint arXiv:2112.08333*, 2021.

[77] Sen Huang, Kaixiang Yang, Sheng Qi, and Rui Wang. When large language model meets optimization, 2024.

[78] Seokhwan Kim, Mihail Eric, Karthik Gopalakrishnan, Behnam Hedayatnia, Yang Liu, and Dilek Hakkani-Tur. Beyond domain apis: Task-oriented conversational modeling with unstructured knowledge access, 2020.

[79] Traian Rebedea, Razvan Dinu, Makesh Sreedhar, Christopher Parisien, and Jonathan Cohen. Nemo guardrails: A toolkit for controllable and safe llm applications with programmable rails, 2023.

[80] Junfeng Jiao, Saleh Afroogh, Yiming Xu, and Connor Phillips. Navigating llm ethics: Advancements, challenges, and future directions, 2024.

[81] Lochan Basyal and Mihir Sanghvi. Text summarization using large language models: A comparative study of mpt-7b-instruct, falcon-7b-instruct, and openai chat-gpt models, 2023.

[82] Jiahuan Pei, Pengjie Ren, and Maarten de Rijke. A modular task-oriented dialogue system using a neural mixture-of-experts, 2019.

[83] Q. Vera Liao and Jennifer Wortman Vaughan. Ai transparency in the age of llms: A human-centered research roadmap, 2023.

[84] Adarsh MS, Jithin VG, and Ditto PS. Efficient hybrid inference for llms: Reward-based token modelling with selective cloud assistance, 2024.

[85] Nalin Kumar and Ondřej Dušek. Leeets-dial: Linguistic entrainment in end-to-end task-oriented dialogue systems, 2024.

[86] Ziyang Chen and Stylios Moscholios. Using prompts to guide large language models in imitating a real person's language style, 2024.

[87] Hao Sun, Xiao Liu, Yeyun Gong, Yan Zhang, Daxin Jiang, Linjun Yang, and Nan Duan. Allies: Prompting large language model with beam search, 2023.

[88] Ivan Sekulić, Silvia Terragni, Victor Guimarães, Nghia Khau, Bruna Guedes, Modestas Filipavicius, André Ferreira Manso, and Roland Mathis. Reliable llm-based user simulator for task-oriented dialogue systems, 2024.

[89] Zelong Li, Shuyuan Xu, Kai Mei, Wenyue Hua, Balaji Rama, Om Raheja, Hao Wang, He Zhu, and Yongfeng Zhang. Autoflow: Automated workflow generation for large language model agents, 2024.

[90] Nikolay Bogoychev, Pinzhen Chen, Barry Haddow, and Alexandra Birch. The ups and downs of large language model inference with vocabulary trimming by language heuristics, 2024.

[91] Wangsong Yin, Mengwei Xu, Yuanchun Li, and Xuanzhe Liu. Llm as a system service on mobile devices. *arXiv preprint arXiv:2403.11805*, 2024.

[92] Tianyu Cui, Shiyu Ma, Ziang Chen, Tong Xiao, Shimin Tao, Yilun Liu, Shenglin Zhang, Duoming Lin, Changchang Liu, Yuzhe Cai, Weibin Meng, Yongqian Sun, and Dan Pei. Logeval: A comprehensive benchmark suite for large language models in log analysis, 2024.

[93] Lu Wang, Max Song, Rezvaneh Rezapour, Bum Chul Kwon, and Jina Huh-Yoo. People's perceptions toward bias and related concepts in large language models: A systematic review, 2024.

[94] Da Song, Xuan Xie, Jiayang Song, Derui Zhu, Yuheng Huang, Felix Juefei-Xu, and Lei Ma. Luna: A model-based universal analysis framework for large language models, 2024.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.