A Comprehensive Survey on Large Language Models for Task-Oriented Dialogue Systems

SurveyForge

Abstract— Large language models (LLMs) have significantly advanced task-oriented dialogue systems by enabling contextually rich and flexible interactions. This survey highlights key research dimensions, including architectural foundations, pre-training and fine-tuning methodologies, memory and context management, and training paradigms, showcasing their role in enhancing task completion rates, user satisfaction, and adaptability. It discusses the integration of external knowledge bases, multimodal inputs, and real-time context refinement as pivotal trends to address challenges such as response hallucinations, error propagation, and domain adaptation. Emerging approaches like reinforcement learning from human feedback, retrieval-augmented generation, and adaptive learning mechanisms have propelled personalization and robustness while maintaining contextual accuracy. However, issues like ethical biases, computational costs, and interpretability persist, limiting deployment and scalability. Additionally, evaluation frameworks combining quantitative metrics with user-centered assessments are identified as critical gaps in current methodologies. The survey underscores future directions in multimodal systems, continuous learning paradigms, and hybrid architectures, alongside a call for transparent and standardized evaluation frameworks. By fostering interdisciplinary collaboration and responsible AI practices, LLM-based dialogue systems hold immense potential for transforming interactions across diverse applications, from customer service to healthcare, while aligning with societal and ethical expectations.

Index Terms—large language models, task-oriented dialogues, adaptive learning mechanisms

1 Introduction

Task-oriented dialogue systems (TODs) represent a crucial intersection of artificial intelligence (AI), natural language processing (NLP), and user experience, facilitating automated interactions in diverse applications such as customer service, personal assistants, and information retrieval. Their significance has burgeoned in recent years, driven by a growing reliance on automated solutions to enhance user engagement and operational efficiency. The evolution of these systems has been marked by a transition from traditional rule-based paradigms to sophisticated models powered by large language models (LLMs), which have demonstrated a compelling ability to understand and generate natural language, thereby revolutionizing the capabilities of dialogue systems.

Historically, dialogue systems have progressed through distinct stages, from simple scripted interactions constrained by pre-defined rules to modular systems that manage various dialogue stages through explicit task decomposition. This approach, although structured, often suffered from limitations such as error propagation across modules and inflexible responses to user inputs. Recent innovations in deep learning, particularly with the introduction of models like transformers, have enabled more coherent and contextually aware dialogue interactions. The embrace of LLMs, exemplified by architectures such as GPT-3 and its successors, marks a transformative leap, allowing for the generation of nuanced, contextually relevant responses that enhance the overall user experience [1].

The integration of LLMs into task-oriented dialogue systems has catalyzed significant advancements in key performance metrics, including task completion rates, user satisfaction, and system robustness. Compared to earlier methodologies, LLMs exhibit the ability to engage in open-ended conversations while still delivering on the specific goals of the task. For instance, recent studies have demonstrated that incorporating user feedback mechanisms within LLMs can drastically improve responsiveness and adaptiveness [2]. Moreover, these models can effectively manage the dual challenge of understanding user intents while generating accurate replies, which is vital for applications requiring a high degree of specificity, such as banking or e-commerce [3].

However, the deployment of LLMs in TODs is not without challenges. One major concern is the phenomenon known as hallucination, where models generate inaccurate or misleading information that can mislead users and erode trust. This issue demands robust mitigation strategies, such as leveraging external knowledge bases to ground dialogue responses or employing reinforcement learning from human feedback to refine model outputs [4]. Furthermore, biases inherent in the training data pose ethical dilemmas that necessitate careful consideration and proactive intervention, ensuring that automated systems do not perpetuate societal inequalities [5].

Emerging trends indicate a shift towards leveraging multimodal inputs, where dialogue systems not only process text but also incorporate visual and auditory information, enhancing contextual understanding and interactivity [6]. This trend, coupled with advancements in adaptive learning algorithms that allow systems to personalize user interactions based on historical data, is indicative of a future where systems become increasingly attuned to individual user needs.

As we chart the path forward, the continued evolution of LLMs presents exciting possibilities for enhancing task-oriented dialogue systems. Future research must focus on refining evaluation methodologies, addressing the balance between model complexity and accessibility, and establishing frameworks for responsible AI deployment. By fostering interdisciplinary collaboration among AI, ethics, and user experience design, the dialogue systems of tomorrow have the potential to offer not only functional efficacy but also a rich and engaging user-centric experience in diverse conversational landscapes.

2 Architectural Foundations of Large Language Models

2.1 Transformer Architecture and Self-Attention Mechanism

The transformer architecture, first introduced by Vaswani et al., has fundamentally transformed the landscape of neural network design, particularly for large language models (LLMs). At the core of this architecture is the self-attention mechanism, which enables the model to weigh the importance of different words irrespective of their positions in a sequence. This capability is particularly crucial for dialogue systems, where understanding context and inferring relationships between conversational components are essential for effective interaction and task completion.

Self-attention operates by creating a set of learned representations in which each word in the input sequence interacts with every other word. This interaction is quantified through three primary vectors: the query (Q), the key (K), and the value (V). For a given input word, the self-attention score is computed as a dot product of its query with all the keys, normalized using a softmax function to produce attention weights:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where d_k is the dimension of the keys. This equation signifies how much focus should be placed on each word when constructing the representation for a particular input, enabling the model to capture contextual dependencies effectively. Notably, this architecture allows for parallelization during training, leading to significant improvements in computational efficiency compared to recurrent neural networks (RNNs), which process sequences in a strictly linear manner. This attribute is especially advantageous for dialogue systems that often require handling long contexts and multiple turns in conversation, where sequential processing would be prohibitively slow.

Further enhancing the capabilities of self-attention is the multi-head attention mechanism. By allowing the model to jointly attend to information from different representation subspaces, multiple attention heads can learn various aspects of context. This holistic approach not only enriches the contextual understanding but also supports the model's ability to generate diverse and nuanced responses in dialogue settings. Lin et al. [7] emphasize that such multi-dimensional representation learning can lead to more engaging and coherent conversation flows.

However, the transformer architecture is not without its limitations. One significant challenge lies in its tendency to yield overly attention-focused outputs that may lead to incoherent or irrelevant responses—often referred to as hallucinations. This phenomenon arises from the model's intrinsic reliance on previously generated tokens and the complexities involved in maintaining context across longer dialogues. Recent research illustrates that this issue persists even in advanced architectures that employ transformers, necessitating further innovations in dialogue management strategies [8], [9].

Another trade-off inherent in the transformer architecture is the computational cost associated with the self-attention mechanism, especially as the input size increases. The complexity of self-attention is $O(n^2)$ with respect to the sequence length n, which can become unmanageable in resource-constrained environments [1], [10]. This has sparked research into efficiency improvements, such as sparse attention mechanisms and reduced-rank approximations.

Emerging trends point toward the integration of self-attention with other models and methodologies. In particular, the incorporation of memory-augmented networks and retrieval-augmented generation (RAG) approaches offer promising pathways for enhancing the contextuality of dialogue responses without sacrificing the efficiency gains of transformers [4], [11]. Furthermore, understanding and mitigating issues related to bias and interpretability continue to present both challenges and opportunities for future research in the realm of LLMs, particularly in sensitive applications such as task-oriented dialogues [12].

In summary, while the transformer architecture and its self-attention mechanism represent a significant leap forward in natural language processing capabilities, ongoing research must address its challenges and limitations. Future directions may explore hybrid models that combine the strengths of transformers with other learning paradigms to create more robust and contextually aware dialogue systems. Integrating external knowledge bases and enhancing dynamic memory management will be key to developing models that can sustain coherent, human-like interactions over extended conversational periods.

2.2 Pre-Training and Fine-Tuning Methodologies

Pre-training and fine-tuning methodologies are essential components in developing large language models (LLMs) for task-oriented dialogue systems. Pre-training involves exposing models to vast datasets to learn generic language representations, while fine-tuning tailors these representations for specific dialogue tasks, effectively bridging the gap between broad language understanding and task-focused performance.

In the pre-training phase, substantial corpora, often comprising diverse texts across different domains, are utilized, allowing models to capture the broad syntactic structures and semantic relationships inherent to the language. This phase employs unsupervised learning techniques with objectives such as masked language modeling (MLM) or next sentence prediction, exemplified by architectures like BERT and its variants. For instance, BERT's MLM task enhances the model's contextual understanding by predicting masked words within sentences, thereby improving its ability to infer meaning through context [13]. The resultant embeddings create a foundational language model with a generalized understanding of the linguistic landscape.

Transitioning to fine-tuning, this stage adapts the robust foundational models for specific tasks through supervised learning on smaller, domain-specific datasets, which contain labeled examples outlining goals, intents, and user expectations. Fine-tuning techniques can vary; standard approaches often leverage cross-entropy loss for classification tasks within dialogue systems. This process allows the model to refine its understanding of task-specific nuances and dialogue behaviors [14]. However, fine-tuning can be resource-intensive, necessitating careful oversight to avoid overfitting, particularly when the fine-tuning dataset is limited in size.

To enhance the efficacy and efficiency of fine-tuning, a variety of techniques have emerged. Parameter-efficient adaptations such as adapters or low-rank adaptations allow for significant reductions in the number of trainable parameters, thus minimizing computational costs while still bolstering model performance [15]. Such strategies stand in contrast to full fine-tuning, offering pathways for adapting models with minimal retraining complexity—an essential advantage for smaller organizations with constrained computational resources.

The utilization of transfer learning represents another critical advancement, where knowledge from previously fine-tuned models is leveraged for new tasks [16]. This methodology accelerates the learning process for new tasks by initializing the model with weights adapted from closely related tasks, subsequently promoting faster convergence and improved performance, even when the number of annotated examples is limited.

Recent studies have brought attention to few-shot and zero-shot learning paradigms, which enable LLMs to generalize to unseen tasks with minimal examples or direct instructions [17]. These innovative approaches drastically reduce the data requirements typically necessary for effective fine-tuning, revealing a trend towards maximizing model versatility while minimizing resource consumption.

Despite these advancements, several challenges persist. Models often exhibit domain adaptation issues, where performance can significantly degrade in novel settings, underscoring the need for robust techniques in this area. Furthermore, enhancing interpretability and transparency in how fine-tuned models derive decisions remains critical for building user trust in automated dialogue systems [18].

In synthesis, the interplay between pre-training and fine-tuning methodologies continues to evolve, driven by the increasing need for efficient, adaptable, and transparent dialogue systems. As LLMs advance in their capabilities for more complex language tasks, ongoing research is vital to address current limitations and explore innovative techniques that optimize their applications in real-world scenarios. Addressing challenges in interpretability, efficiency, and domain adaptation is likely to enhance user experiences and broaden the applicability of LLMs across various fields.

2.3 Memory and Context Management Techniques

Memory and context management techniques play a pivotal role in enhancing the performance of large language models (LLMs) in task-oriented dialogue systems. These systems face the challenge of maintaining coherence and relevance across multi-turn interactions, which is crucial for user satisfaction and effective task completion. The strategies employed for this purpose can be broadly categorized into memory-augmented approaches, context encoding mechanisms, and dynamic context refinement.

Memory-augmented frameworks leverage external memory structures to persist information across dialogue turns. For instance, memory networks, which allow the model to read from and write to an external memory, can significantly enhance context retention. These models can store user-specific interactions and recall them as needed, thereby personalizing dialogues and improving continuity. Studies like those described in [19] illustrate how integrating user history into dialogue processes leads to higher engagement and satisfaction rates. However, while memory-augmented approaches provide a robust means for managing longer dialogues, they do introduce complexities related to memory management, such as deciding what to store, when to retrieve, and how to update the memory efficiently.

Another pivotal technique involves the implementation of context encoding via mechanisms like attention. The transformer architecture employs a self-attention mechanism that allows models to weigh the relevance of previous dialogue turns dynamically. This capability is particularly useful in task-oriented systems where users may refer back to earlier elements in the conversation. The multi-head attention formulation enables LLMs to simultaneously consider different parts of the prior context, as highlighted in [20]. This approach enhances the model's ability to capture nuances and maintain engagement over extended dialogues. However, a potential challenge lies in the computational cost of processing long sequences, which can limit the effective context length that these models can handle efficiently.

Dynamic context management is another emerging strategy, which emphasizes selective retrieval and relevance-based referencing of context. This approach continuously updates the context based on user interactions and the dialogue state, allowing models to prioritize the most relevant historical information while minimizing noise from less pertinent dialogue turns. Techniques such as adaptive context windows, which discard older turns that have less impact on the current interaction, have shown promise in maintaining conversational fluidity and coherence. For example, work on transfer learning applied to context-aware dialogue systems indicates that such dynamic approaches can improve user experience while optimizing resource utilization by focusing computational power on the most relevant context, as seen in models described in [3].

Evaluating these context management techniques reveals trade-offs between memory depth, computational efficiency, and dialogue relevance. While memory-augmented approaches enhance depth, the risk of overfitting to historical data and the potential for information overload must be managed carefully. On the other hand, approaches that streamline context retrieval prioritize efficiency but may sacrifice richness in nuanced dialogue understanding. Future research should therefore explore hybrid approaches that combine these strategies to balance depth and responsiveness.

As LLMs continue to evolve, integrating multimodal inputs into dialogue systems further complicates memory and context management. Techniques that enable models to hybridize text with visual or auditory data could allow for even richer contextual understanding and management, fostering deeper engagement. In conclusion, advancing memory and context management techniques remains a crucial area of research within task-oriented dialogue systems, necessitating innovative solutions that ensure coherent, contextually aware, and user-centric interactions. Ongoing efforts should concentrate on developing adaptive strategies that efficiently leverage both memory and context, ultimately enhancing the effectiveness of LLMs in practical applications.

2.4 Enhancements in Training Paradigms

Recent advancements in training paradigms have significantly enhanced the efficacy of large language models (LLMs) for task-oriented dialogue systems. This subsection explores cutting-edge techniques such as reinforcement learning from human feedback (RLHF), curriculum learning, and multi-task learning, which work in concert to fortify the adaptability and functionality of models within dynamic conversational environments.

A cornerstone of these advancements is RLHF, which enables LLMs to align more closely with user preferences by learning from explicit human feedback. In conventional supervised learning, models are trained on fixed datasets, which can produce suboptimal performance when user preferences deviate from those represented in the training data. RLHF addresses this limitation by establishing iterative feedback loops, wherein models generate responses that are subsequently rewarded or penalized based on user satisfaction metrics. The integration of RLHF has been shown to enhance user engagement and satisfaction in dialogue systems by fine-tuning responses to better meet human expectations, as demonstrated in studies related to dialog control optimization using RL techniques [14]. However, a notable trade-off associated with RLHF is the computational cost and time required for iterative training, particularly over large datasets, which can hinder rapid deployment in resource-constrained environments.

Another important training paradigm is curriculum learning, which enhances training efficiency by progressively structuring the learning tasks. With this approach, models initially tackle simpler tasks, gradually advancing to more complex ones. This structured progression enables smoother learning curves and improved model stability. By introducing increasingly challenging dialogue scenarios, models can expand their competence incrementally, thereby boosting their utility in real-world applications where conversational complexity often varies [21]. The effectiveness of curriculum learning lies in its ability to mitigate overfitting on simplified data, although it requires meticulous planning of the training regimen, adding complexity compared to traditional methods.

Multi-task learning is also gaining recognition as a robust strategy for leveraging shared knowledge across various tasks, thereby enhancing model generalization. By simultaneously training on related dialogue tasks, knowledge acquired in one domain can benefit others, promoting efficient learning and reducing the overall data volume required [22]. This approach is especially pertinent in dialogue systems where user intents may shift rapidly, necessitating a flexible model capable of adapting to contextual changes. However, one challenge to consider is the potential for negative transfer, whereby learning tasks may interfere with one another, potentially degrading performance in specific scenarios.

Emerging hybrid approaches that synergize these methodologies are gaining traction. For instance, combining RLHF with curriculum and multi-task learning may yield systems that not only quickly adapt to user feedback but also seamlessly transition between diverse dialogue contexts. This confluence of strategies supports the development of more resilient models capable of maintaining coherent and contextually appropriate dialogues, even in multi-turn scenarios where memory management is critical [23], [24].

The practical implications of these advancements are substantial, particularly in rapidly evolving sectors such as customer service and healthcare, where user expectations and conversational complexities can fluctuate widely. As dialogue systems progress, future training paradigms are likely to incorporate even more sophisticated forms of dynamic adaptation, potentially leveraging real-time user interactions to continually refine and adjust model responses through online learning techniques.

In conclusion, the integration of advanced training paradigms like RLHF, curriculum learning, and multi-task learning represents a pivotal advancement in enhancing the performance of large language models for task-oriented dialogue systems. As these techniques develop and converge, the focus will likely shift toward creating efficient training methodologies that balance computational efficiency with model adaptability, ensuring these systems can fulfill the demands of users in fast-paced environments. The ongoing exploration of these paradigms remains vital in the evolution of intelligent conversational agents, making significant strides in user-centric design and functionality.

2.5 Architectural Extensions and Variants

The evolution of transformer architectures has paved the way for various architectural extensions and variants that significantly enhance the capabilities of large language models (LLMs) in managing dialogues effectively. At the core of these advancements lies a need to address the inherent limitations of foundational models in terms of dialogue coherence, contextual understanding, and real-time adaptability. This subsection explores notable extensions and their ramifications on task-oriented dialogue systems.

One of the primary architectural innovations is the adoption of decoder-only and encoder-decoder models tailored for specific dialogue requirements. Decoder-only models like GPT leverage unidirectional generation, which facilitates fluency in text generation but often struggles with context management in multi-turn dialogues. Conversely, encoder-decoder architectures, such as T5 and BART, are adept at handling input context from both user queries and internal states, thereby improving coherent response generation in task-oriented scenarios. Studies have demonstrated that encoder-decoder models can better capture the intricacies of dialogue context, as they can learn to attend to both current inputs and previous dialogue history [18].

Integrating memory and reasoning modules into traditional architectures has also shown promising results. For instance, models augmented with attention mechanisms that specifically target episodic memory can retain relevant information from past interactions, allowing for more contextually aware responses. This architecture facilitates dialogue continuity over extended interactions, addressing a significant limitation wherein standard models often forget crucial prior context [25]. The introduction of structured memory, represented as an external Knowledge Graph, enables LLMs to access factual information more reliably, complementing their generative capabilities and reducing hallucinations [4].

Another noteworthy trend is the development of retrieval-augmented generation (RAG) approaches, where LLMs are combined with external retrieval mechanisms to access pertinent knowledge bases during generation. This hybridization enables the model to ground its responses in factual data, which is particularly beneficial in task-oriented dialogues that demand accurate and relevant information [26]. RAG has been empirically validated to enhance the relevance and correctness of dialogue system outputs, thus boosting user satisfaction and trust [4].

Moreover, recent advancements in multimodal integration have begun to reshape the architecture paradigms of LLMs. By employing models that process both text and visual inputs, systems can achieve a more nuanced understanding of user intentions and contexts. This integration is particularly useful in applications proposing complex queries that benefit from visual data, as highlighted in [27]. As systems progress towards holistic modalities, they create richer interaction experiences, enhancing the effectiveness of task-oriented dialogues.

Future directions appear promising as these architectural extensions challenge traditional paradigms by addressing the scalability and adaptability of LLMs. However, challenges remain in balancing model complexity with efficient training and inference mechanisms. Emerging techniques such as model-adaptive prompt optimization and dynamic task-related knowledge retrieval show great potential for optimizing performance while maintaining manageable resource requirements [28], [29].

In conclusion, the ongoing evolution of transformer architectures through targeted extensions and adaptations is critical for advancing the performance of dialogue systems. As LLMs continue to incorporate memory, reasoning, and multimodal capabilities, they will likely redefine how users interact with AI in task-oriented scenarios, pushing the boundaries of what these systems can achieve in real-world applications.

3 Adaptation Techniques for Task-Oriented Dialogue Systems

3.1 Prompt Engineering Strategies

Prompt engineering has emerged as a critical strategy in adapting large language models (LLMs) for specific task-oriented dialogue applications. By designing effective prompts, developers can significantly enhance the performance of LLMs, guiding them toward producing contextually relevant and coherent responses. This subsection systematically explores various prompt engineering strategies, elucidating their methodologies, strengths, limitations, and implications for the design of task-oriented dialogue systems.

One primary approach in prompt engineering is **hard prompting**, wherein explicit, fixed prompts are crafted to yield precise instructions. Such prompts provide direction on the desired outputs, which often leads to improved model coherence and performance. For example, a prompt structured to request specific factual information will typically generate more accurate responses than an ambiguous one. This approach has been corroborated in studies demonstrating that models trained with explicit prompts outperform those relying on broader or more generalized input formats [30]. However, hard prompting may hinder flexibility, as changes to the tasks could necessitate substantial prompt reconfigurations.

Conversely, **soft prompting** employs continuous representations of prompts that allow for a more fluid interaction with LLMs, often employing embedding techniques to adapt prompts dynamically according to specific dialogue contexts. This approach is particularly beneficial in maintaining a conversational flow, as seen in systems that adjust the tone or complexity of language based on user inputs and historical dialogue turns [31]. Although soft prompting enhances adaptability, it requires careful tuning to ensure that the underlying model adequately recognizes and leverages context throughout the dialogue.

Another innovative strategy gaining traction is **automatic prompt generation**, wherein LLMs are employed to generate and optimize prompts themselves. Techniques like Automatic Prompt Engineering (APE) have been proposed, which use empirical performance data to refine prompts iteratively. This automation alleviates the labor-intensive nature of prompt design and allows for rapid prototyping and adjustment according to performance feedback [28]. While promising, the challenge lies in ensuring that the generated prompts retain fidelity to user goals and do not introduce unintended biases or inaccuracies.

A fusion of hard and soft prompt techniques, such as utilizing a sparse set of examples in a few-shot learning setup, has shown great promise. In such scenarios, the model is exposed to a limited number of context-specific examples to guide the generation process. This method leverages the strengths of in-context learning while maintaining a degree of control over the prompt structure, allowing for prompt flexibility without sacrificing specificity [32]. Nevertheless, a limitation to note is the model's reliance on the quality and representativeness of the examples provided—a challenge that requires meticulously curated datasets.

Emerging trends in prompt engineering highlight the burgeoning interest in techniques that incorporate multimodal data, integrating visual or audio input to enrich the prompt context and enhance dialogue versatility. Such advancements are pivotal in scenarios requiring richer context beyond textual information, thereby potentially dominating the next generation of task-oriented dialogue systems [33]. However, these integrations also spur additional complexities, necessitating robust frameworks to manage multimodal inputs effectively.

In conclusion, prompt engineering remains a pivotal area of exploration for tailoring LLMs to task-oriented dialogue systems, with various methodologies that offer distinct advantages and drawbacks. As future research delves into hybrid approaches, incorporating multimodal and automatic generation strategies, the potential for enhancing conversational quality and user satisfaction in dialogue systems expands significantly. Addressing the challenges associated with biases and adaptability will be paramount in realizing the full potential of prompt-driven dialogue systems, thereby ensuring alignment with user expectations and diverse application scenarios.

3.2 Transfer Learning Techniques

Transfer learning has emerged as a pivotal methodology for fine-tuning pre-trained language models (PLMs) within task-oriented dialogue systems, facilitating rapid adaptation to specific domains with significantly reduced data requirements. This approach capitalizes on the vast linguistic knowledge embedded in these models, acquired during extensive pre-training on diverse textual datasets. In this subsection, we will analyze several prominent techniques in transfer learning applicable to dialogue systems, focusing on fine-tuning strategies, parameter-efficient methods, and domain adaptation approaches, alongside their respective strengths and limitations.

The most commonly employed strategy in transfer learning is **fine-tuning**, where a pre-trained model is updated on a smaller, task-specific dataset. This approach optimizes all model parameters across a limited number of labeled examples, effectively aligning the model's responses with the desired output of a specific task. An example of this can be seen in the end-to-end LSTM-based dialogue control system, which utilizes both supervised learning (SL) and reinforcement learning (RL) to enhance performance by leveraging predefined exemplars of dialogues. This demonstrates the benefits of combining learning paradigms [14]. While fine-tuning can yield high performance on specific tasks, it faces challenges related to data scarcity and the risk of overfitting, particularly in low-resource settings.

In response to the increasing demands for computational efficiency and flexibility, **parameter-efficient transfer learning** techniques have gained traction. Approaches such as adapters or low-rank adaptation modify only a small subset of parameters, allowing for rapid adaptation without retraining the entire model. An illustrative case is the Adapter-Bot, which capitalizes on fixed backbone models and triggers various dialogue skills via lightweight adapters, thereby substantially reducing the computational overhead associated with full model fine-tuning [34]. This method enables diverse knowledge integration while maintaining adaptability across multiple tasks. However, while these adaptations mitigate the risks of overfitting, they may still underperform compared to fully fine-tuned models in high-stakes applications that require nuanced understanding.

Another significant area within transfer learning is **domain adaptation**, which harnesses knowledge from related domains to bolster performance in under-resourced areas. Such approaches have shown efficacy in leveraging existing user interactions to supplement training, particularly in domains where annotated data is scarce. Techniques like bootstrapping from large dialogue corpora have proven beneficial in enhancing dialogue systems' understanding of various conversational dynamics and contexts [16]. Furthermore, multi-task learning frameworks can be employed to optimize models across varying tasks simultaneously, leveraging synergies between them to improve overall performance [35].

Emerging trends in transfer learning now emphasize the use of auxiliary self-supervised tasks that augment traditional supervised training approaches, enabling models to derive richer contextual features from dialogue data. For instance, models that learn alongside tasks such as next utterance prediction or dialogue coherence recognition can refine the language model's understanding of conversation flow without extensive labeled data [36]. These innovations not only enhance adaptability but also lead to robustness in increasingly complex interactive environments.

Despite these advancements, several challenges remain in the effective application of transfer learning within task-oriented dialogue systems. The need for extensive labeled datasets persists, particularly when aiming for high accuracy in multi-turn dialogue contexts. Additionally, integrating contextual memory—allowing models to maintain a coherent narrative across lengthy interactions—represents a formidable technical hurdle as existing model architectures grapple with retaining relevant information over time. Addressing these issues necessitates the development of novel architectures capable of efficiently managing both the dynamic aspects of dialogue and the static knowledge inherent in large models.

In conclusion, the evolution of transfer learning techniques within task-oriented dialogue systems showcases a landscape rich with innovation and potential. By continuously refining adaptation strategies, dialogue systems can achieve enhanced performance across diverse domains while remaining agile in meeting the increasingly complex demands of user interactions. Future research directions should focus on overcoming current limitations through interdisciplinary approaches that blend insights from cognitive sciences and computational linguistics, ultimately leading to more intuitive and effective dialogue agents.

3.3 Few-Shot and Zero-Shot Learning

Large language models (LLMs) have demonstrated remarkable capabilities in few-shot and zero-shot learning, which are crucial for task-oriented dialogue systems, especially when domain-specific training data is limited. Few-shot learning allows models to perform tasks after being given only a handful of examples, while zero-shot learning enables models to generalize to new tasks without any specific examples, instead relying on their pre-existing knowledge and contextual cues. These mechanisms are particularly valuable in real-world applications of dialogue systems, where data collection can be resource-intensive and not always feasible.

In few-shot learning, LLMs utilize in-context learning techniques, where the model is primed with task-specific examples provided within the input context. This approach allows the model to adapt its responses based on the examples given. Research has shown that models like GPT-3 can achieve reliable performance on unseen tasks by leveraging a few exemplars, showcasing their ability to generalize effectively. Studies such as [37] highlight the effectiveness of this strategy, demonstrating that such models can adapt their understanding of Natural Language Understanding (NLU), Dialogue State Tracking (DST), Dialogue Policy (DP), and Natural Language Generation (NLG) tasks from minimal input. However, the accuracy of predictions can be sensitive to how the few examples are structured and the specific wording used, indicating a potential trade-off between flexibility and performance.

Conversely, zero-shot learning leverages the intrinsic properties of large pre-trained models, allowing them to tackle tasks for which they have not been explicitly trained. This ability arises from the extensive training these models undergo on a diverse array of tasks and domains, enabling them to infer and generate relevant outputs based on context alone. The findings from [38] emphasize that LLMs can effectively bypass traditional training pipelines by using structured prompts, guiding the models to navigate complex queries successfully even in the absence of task-specific examples.

Moreover, several innovations have emerged to enhance the performance of LLMs in zero-shot settings. For instance, prompt engineering plays a critical role in framing questions and tasks to minimize ambiguity and maximize the usefulness of the contextual information provided. Effective prompt design can significantly influence the model's understanding and response generation, illustrating the interplay between user inquiries and model outputs. Techniques such as multi-stage prompting, as explored in [39], further refine how knowledge is extracted and responses are generated, augmenting zero-shot capabilities by focusing on knowledge relevance alongside response generation.

Despite their advantages, few-shot and zero-shot learning approaches in LLMs face challenges. Model performance can fluctuate based on the complexity and novelty of the task, as well as the intricacies of the prompts. Furthermore, LLMs are still prone to issues such as hallucinations—generating incorrect or nonsensical information—which can compromise the reliability of dialogue systems. Techniques to mitigate these risks typically involve a combination of robust prompt design, augmentation with relevant knowledge bases, and user feedback to improve iterative learning mechanisms [15].

Looking toward future developments, the integration of few-shot and zero-shot learning capabilities into task-oriented dialogue systems will likely depend on the continuous enhancement of model architectures and training paradigms. Emerging trends indicate that combining these models with retrieval-based frameworks could enhance context comprehension and response accuracy in dynamic interactions, paving the way for more intelligent and adaptable dialogue systems. Continued exploration of these adaptive strategies will foster more resilient dialogue agents capable of addressing nuanced user needs across various domains, ultimately improving the efficiency and personalization of customer interactions in task-oriented scenarios.

3.4 User-centric Adaptation Techniques

User-centric adaptation techniques are vital for enhancing task-oriented dialogue systems' interactions by personalizing responses based on individual user preferences and requirements. These approaches leverage user data, interaction history, and engagement metrics to create more relevant and responsive dialogue experiences. By emphasizing the need to tailor dialogue systems for individual user needs, we can significantly improve engagement and task satisfaction, ultimately leading to more effective automated interactions.

A prevalent method for user-centric adaptation is personalization, where models learn from historical interactions to customize responses that align closely with user expectations. Recent work showcases how robust user profiles can be developed through continual learning from ongoing interactions, allowing dialogue systems to adapt to varying user personalities and preferences. This personalization fundamentally alters response strategies to ensure contextual relevance and user satisfaction. For instance, methods described in [40] integrate personalization to enhance user engagement by refining models based on individual feedback, ensuring responses are not only contextually aligned but tailored to user profiles as well.

Contextual awareness is another essential aspect of user-centric adaptation, where dialogue systems enhance their ability to retrieve and utilize relevant contextual information across multi-turn conversations. Implementing effective memory management techniques, such as those explored in [23], enables these systems to dynamically maintain and reference previous interactions. This enhances conversational coherence, allowing for heightened relevance in responses by recalling major themes and user intents. The integration of stateful memory structures not only helps preserve context but also facilitates more informed responses throughout extended dialogues.

An innovative trend in user-centric adaptation involves simulating user interactions through LLMs. By generating hypothetical user dialogues, as demonstrated in [41], researchers can create extensive training datasets that reflect real-world conversational dynamics. This method effectively alleviates the reliance on annotated human dialogue data, which is often scarce and costly to compile. The adaptability of these models ensures they cater to diverse user needs, accommodating a wide range of dialogue scenarios without requiring exhaustive annotated corpora.

Evaluating the strengths of user-centric adaptation techniques reveals distinct trade-offs. While personalization undoubtedly boosts user engagement and satisfaction, it also necessitates careful handling of user data to prevent privacy violations and potential biases in model responses. Techniques that rely on context, such as memory-augmented systems, can introduce complexity in dialogue management, demanding more sophisticated algorithms to balance relevance and coherence without overwhelming users with redundant information. Furthermore, simulated user interactions present a scalable alternative for training but could lead to overfitting if not managed correctly, as these synthetic dialogues may not capture the full variability present in real conversations.

Ongoing research continues to address these challenges by exploring advanced methodologies such as reinforcement learning frameworks, which enable dialogue systems to refine their responses based on actual user feedback (as evidenced by strategies in [14]). The incorporation of mechanisms for ethical adaptability—considering user feedback while ensuring equitable treatment across diverse user groups—is also gaining traction. As dialogue systems evolve, fostering greater user trust and understanding through transparent personalization strategies becomes paramount.

In conclusion, user-centric adaptation techniques present a compelling opportunity to significantly enhance task-oriented dialogue systems. Looking ahead, integrating more advanced machine learning techniques, data privacy considerations, and ethical frameworks will be critical. These elements will shape the next generation of intelligent dialogue systems that not only respond to user queries but engage meaningfully with their users, ensuring that technology aligns with individual needs and experiences.

3.5 Hybrid Methods of Adaptation

Hybrid methods of adaptation represent a strategic confluence of diverse approaches aimed at enhancing the performance of task-oriented dialogue systems powered by large language models (LLMs). These systems, characterized by their reliance on data-driven algorithms, benefit from combining techniques such as prompt engineering, transfer learning, and reinforcement learning to achieve robust adaptability across varied tasks and contexts.

One notable aspect of hybrid adaptation is the integration of structured prompt engineering with fine-tuning based on domain-specific datasets. In scenarios where conversational agents must navigate complex user intents, sophisticated prompt configurations can set the stage for effective communication. By leveraging well-crafted prompts that reflect task-specific nuances, LLMs are guided toward generating contextually appropriate responses. For instance, the incorporation of task-instruction templates within prompts often leads to improved understanding of user queries, as seen in the work of Instruction Tuning for Large Language Models [42]. Moreover, combining these structured prompts with rich datasets for fine-tuning allows for robust model adaptability best exemplified in the domain-adaptive pre-training outlined by DialoGLUE [43].

Additionally, hybrid methodologies often employ reinforcement learning (RL) frameworks in conjunction with imitation learning strategies. Here, dialogue agents can learn from simulated interactions that mirror real-world dynamics, refining their performance through direct, user-guided feedback. This is particularly relevant in developing agents that can efficiently adapt to unstructured domains where annotated data is sparse. The effectiveness of such hybrid models can be seen in the blending of user preferences and reinforcement signals that improve task completion rates, as demonstrated in the studies on reinforcement learning from human feedback [44]. However, a pivotal challenge in such methodologies lies in the transition between phases—namely, from offline training to online interaction—where inconsistencies can emerge if not properly managed.

A compelling area of exploration lies in metadata enhancement and context mapping as part of hybrid adaptation approaches. Utilizing external knowledge bases in conjunction with LLMs provides a structured context that can significantly boost the richness of generated responses. Techniques such as retrieval-augmented generation (RAG) exemplify this hybridization, where LLMs leverage retrieved information from knowledge repositories in real-time during conversation, thereby enriching user interactions with factual accuracy and relevance [4]. This methodology, while powerful, raises concerns regarding system complexity and processing time, demanding a delicate balance between performance and efficiency.

Future research endeavors in hybrid adaptation techniques should aim to identify best practices for integrating various adaptation methods seamlessly. For instance, the interplay between few-shot learning and knowledge retrieval frameworks can be optimized to enhance performance without incurring significant computational overhead. Additionally, as LLMs advance toward greater autonomy and self-adaptation, exploring strategies to enable these models to dynamically adapt to user preferences over time is essential. The integration of longitudinal learning and feedback loops from user interactions can fuel continuous improvement, effectively allowing task-oriented dialogue agents to evolve with changing user needs, a concept emphasized in the notion of continual learning frameworks [25].

In conclusion, hybrid methods of adaptation not only signify an evolution in dialogue systems but also present a fertile ground for future exploration. The integration of multiple approaches fosters a richer landscape for dialogue performance, emphasizing flexibility and robustness. By tackling the challenges of complexity, efficiency, and user experience, researchers can push the boundaries of what is achievable in task-oriented dialogue systems, laying the groundwork for more sophisticated, human-centric conversational agents.

3.6 Evaluation of Adaptation Techniques

The evaluation of adaptation techniques in task-oriented dialogue systems powered by large language models (LLMs) poses a multifaceted challenge, requiring a robust framework that accommodates both quantitative and qualitative metrics. As the application of these models spans various domains, it becomes increasingly imperative to thoroughly assess their adaptation strategies to ensure optimal performance in real-world interactions.

Quantitative performance metrics serve as a foundational pillar for evaluation. Traditional metrics such as accuracy, F1 scores, and precision-recall analyses provide essential insights into model performance. For instance, the dialogue success rate quantitatively measures the ratio of successfully completed dialogue tasks to the total number of interactions, thus illustrating the system's effectiveness in achieving its objectives. Several studies conducted within the context of LLMs indicate that enhancements in task performance often correlate with improvements in these metrics. Fine-tuning strategies that incorporate user-specific data have yielded significant increases in task success rates [15]. However, an overreliance on these quantifiable outcomes can inadvertently overshadow critical factors such as user engagement and satisfaction, emphasizing the necessity for a balanced evaluation strategy.

Complementing quantitative metrics, qualitative user feedback is crucial for capturing the nuanced aspects of human-computer interactions. Surveys focusing on user perceptions related to conversational quality, helpfulness, and overall satisfaction demonstrate a positive correlation with quantitative metrics, highlighting the importance of aligning user expectations with model performance [45]. Tools that effectively harness this qualitative data can inform model adaptations and significantly enhance user experiences. Nonetheless, interpreting qualitative metrics remains a challenge due to their inherently subjective nature, which can vary considerably across different user demographics and contexts.

The integration of benchmark datasets into the evaluation framework facilitates standardized comparisons across various adaptation techniques and their corresponding performance outcomes. Established benchmarks such as MultiWOZ and CamRest676 offer rich resources for dialogue evaluation, enabling researchers to consistently assess the efficacy of LLM adaptations [46]. Comparative studies utilizing these datasets deepen our understanding of how specific techniques, including few-shot learning and reinforcement learning strategies, impact task-oriented dialogues. However, researchers must remain cautious of potential biases within these datasets, which can skew evaluation results if not adequately addressed.

Emerging evaluation techniques, such as automated metrics for dialogue quality assessed by large language models themselves, are gaining traction within the research community. These methodologies aim to reduce reliance on labor-intensive manual evaluations while leveraging the inherent capabilities of LLMs to empirically judge the quality of generated responses. For instance, incorporating reinforcement learning with human feedback can create adaptive systems capable of dynamically adjusting their evaluation criteria based on user interactions, thereby promoting continuous improvement [4].

A critical insight into the evaluation process lies in acknowledging the inherent trade-offs associated with various methods. While automated evaluations may enhance efficiency, they risk overlooking nuanced conversational elements best captured through thorough qualitative assessments. Consequently, a comprehensive evaluation framework that synthesizes both quantitative metrics and qualitative insights is vital for informing model development and deployment strategies. Researchers should also adapt evaluation metrics to address emerging challenges, such as those related to biases in LLMs or their propensity to generate non-factual responses.

In conclusion, the evaluation of adaptation techniques for task-oriented dialogue systems represents a dynamic and evolving area of research, characterized by the interplay between empirical metrics and user-centric assessments. As LLMs continue to advance, ongoing research should focus on refining these evaluation frameworks while tackling new challenges, thereby fostering a comprehensive understanding of both model strengths and improvement areas necessary to effectively meet user expectations. Future work must also explore the design of novel metrics that encapsulate the user experience holistically, paving the way for next-generation adaptation techniques that resonate more profoundly with end-users.

4 Evaluation Methods and Performance Metrics

4.1 Quantitative Evaluation Metrics

Quantitative evaluation metrics play a critical role in assessing the performance of task-oriented dialogue systems powered by large language models (LLMs). These metrics provide a framework for objectively measuring the effectiveness, reliability, and user satisfaction of dialogue interactions. Among the most significant metrics are accuracy, precision, recall, and dialogue success rates, which collectively capture various facets of system performance.

Accuracy is often the foundational metric in evaluating dialogue systems, quantifying the proportion of correct responses among total responses generated. More formally, accuracy can be expressed as:

$$\text{Accuracy} = \frac{N_{correct}}{N_{total}}$$

where $N_{correct}$ is the number of correct responses, and N_{total} represents the total number of responses assessed. While accuracy provides a straightforward insight into performance, it can be misleading in scenarios where the distribution of responses is imbalanced; thus, it is often supplemented with precision and recall metrics, which provide a more nuanced analysis of responses.

Precision, defined as the ratio of relevant responses generated over the total number of responses, elucidates the accuracy of positive predictions made by the system. It can be mathematically expressed as:

$$\text{Precision} = \frac{TP}{TP + FP}$$

where TP denotes true positives and FP denotes false positives. A high precision indicates that most of the generated responses are relevant, but it does not account for missed relevant cases, leading us to consider recall, which measures the system's capability to identify all relevant responses:

$$\text{Recall} = \frac{TP}{TP + FN}$$

where FN represents false negatives. These two metrics are often combined into the F1 score, which is the harmonic mean of precision and recall, offering a balanced view of the system's performance:

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

In the realm of task-oriented dialogues, the dialogue success rate emerges as a critical metric, quantifying the proportion of conversations that successfully achieve their intended objectives. This metric underscores the system's efficacy in real-world applications, particularly when user satisfaction hinges on the system's ability to meet specific user goals. Calculating the dialogue success rate involves tracking the completion of tasks against total interactions:

$$\text{Dialogue Success Rate} = \frac{N_{success}}{N_{total-dialogues}}$$

where $N_{success}$ represents dialogues in which the task was accomplished. Previous research indicates that improving dialogue success rates directly correlates with enhanced user satisfaction and trust in dialogue systems, crucial considerations in practical implementations [7], [47].

Despite their widespread use, these quantitative metrics are not without limitations. For instance, accuracy alone may obscure fine nuances of system performance, especially in highly dynamic conversations where context varies significantly across turns. Recent studies have highlighted the need to develop advanced evaluation frameworks that capture multi-turn dialogue complexities and interdependencies effectively, moving beyond traditional metrics. Emerging metrics, such as engagement scores, which assess the depth of user interaction, and contextual coherence scores, which evaluate the relevance of responses in sustained dialogues, represent promising directions for future research [10], [48].

Innovatively, automated evaluation tools, leveraging LLMs for assessing dialogue quality, are gaining traction. Such methodologies can provide rapid assessments and scalability in the evaluation process, although they bring unique challenges concerning alignment with human judgment. As LLMs become more integrated with dialogue systems, exploring their utility in evaluation will become increasingly important [4], [49].

In conclusion, while traditional quantitative evaluation metrics such as accuracy, precision, recall, and dialogue success rates form the bedrock of evaluating task-oriented dialogue systems powered by LLMs, there is a pressing need to expand the repertoire of metrics to better capture the multifaceted nature of dialogue interactions. Future research should prioritize developing robust, context-aware metrics that reflect the increasingly complex and nuanced capabilities of emerging dialogue technologies, guiding improvements in both evaluation practices and system design. The interplay between quantitative metrics and emerging evaluation methodologies will ultimately shape the future landscape of task-oriented dialogue systems.

4.2 Qualitative User Experience Metrics

Qualitative metrics play a crucial role in evaluating user experiences with task-oriented dialogue systems powered by large language models (LLMs). Unlike quantitative measures, which provide numerical insights into system performance, qualitative metrics delve into subjective dimensions such as user satisfaction, engagement, and emotional resonance during interactions. These factors are paramount for assessing the overall usability and effectiveness of conversational agents, as user experiences directly influence the adoption and long-term success of dialogue systems.

One of the cornerstones for qualitative assessment is user satisfaction surveys, which capture users' perceptions of dialogue quality, helpfulness, and satisfaction levels. These surveys typically employ Likert scales to quantify subjective responses, allowing researchers to analyze trends and correlations. For instance, by correlating satisfaction scores with system performance metrics, developers can identify which aspects of the interaction contribute most significantly to user contentment [50]. However, a notable limitation lies in potential biases driven by users' expectations or the framing of survey questions, which may skew results. To mitigate this, incorporating mixed-method approaches that combine qualitative observations with quantitative data could yield a more nuanced understanding of user experiences.

Engagement metrics serve as another valuable qualitative lens through which to evaluate dialogue systems. Metrics such as conversation length, turn-taking dynamics, and session duration can reveal patterns of user interaction that reflect varying engagement levels. For instance, an increase in the number of turns per session may indicate a more engaging dialogue experience, while excessively long interactions might suggest issues of clarity or relevance in responses. Studies have suggested that fostering an optimal balance in the number of turns can enhance user engagement without leading to fatigue or frustration [51].

An emerging trend in qualitative assessment involves the consideration of emotional valence in dialogue interactions, which quantifies the emotional tone of responses through sentiment analysis. By evaluating the emotional alignment between user inputs and system responses, researchers can gain insights into the empathetic capabilities of dialogue agents [52]. This aspect is particularly relevant in user-oriented scenarios, where emotional resonance can significantly enhance user experience. Nonetheless, challenges persist in robustly defining and measuring emotional constructs in dialogue, given that emotions can be context-dependent and vary widely across user demographics.

Technical advancements in natural language processing, such as hierarchical models or multi-turn response generation strategies, promise to further enhance our understanding of dialogue dynamics by capturing contextual subtleties and facilitating coherent, contextually aware interactions [53]. Future qualitative evaluation methods may integrate real-time feedback loops that enable systems to adjust responses based on ongoing user sentiment and engagement metrics, thereby augmenting system adaptability.

Thus, synthesizing qualitative user experience metrics offers ample opportunities for innovation in dialogue system design. Expanding qualitative metrics to encompass aspects of personalization, such as user profiling and history tracking, could lead to tailored interactions that feel more intuitive and engaging. Moreover, addressing the existing technical limitations in qualitative metrics—such as defining criteria for coherence and relevance—could propel the field toward developing enhanced dialogue systems that not only perform tasks but also foster meaningful interactions. Continued exploration of these dimensions, along with efforts to standardize qualitative evaluation protocols, will be vital in verifying the effectiveness of LLM-based dialogue systems in real-world applications.

4.3 Benchmark Datasets for Evaluation

Benchmark datasets play a pivotal role in the evaluation of task-oriented dialogue systems, serving as crucial resources to establish baselines and compare the performance of various models. These datasets provide standardized tasks that facilitate systematic assessment, enabling researchers to evaluate the effectiveness, scalability, and adaptability of dialogue systems powered by large language models (LLMs). As dialogue systems evolve, it becomes imperative to leverage diverse benchmarks that can adequately represent the complexity and variability inherent in human-computer interactions.

One of the most prominent datasets in this realm is the MultiWOZ dataset, which encompasses dialogues across multiple domains such as hotels, restaurants, and attractions. The dataset facilitates the evaluation of models on multi-turn dialogues, enabling rigorous analysis of dialogue management, dialogue state tracking, and intent recognition. The versatility of MultiWOZ allows for comparison across various methodologies, making it an essential component in studies focused on LLMs [38], [54]. However, despite its comprehensive nature, challenges persist, particularly regarding the dataset's annotation quality and the potential biases inherent in its construction.

Another notable benchmark is the Cambridge Restaurant dataset, specifically designed to evaluate dialogue state tracking within restaurant reservation contexts. This dataset emphasizes task completion and communication efficiency, allowing for a focused examination of dialogue management strategies [3]. While its domain specificity can yield deep insights into restaurant-related dialogues, its applicability is limited when compared to broader datasets like MultiWOZ, which requires models to generalize across various domains and scenarios.

Emerging trends in the creation of synthetic dialogue datasets, such as those developed via Large Language Models, present innovative methods to overcome traditional data scarcity while maintaining quality control. For instance, the PETAL framework leverages transfer learning to optimize personalized dialogue system training using a small set of personalized data alongside a larger collective dataset [19]. This approach highlights the potential for rapidly generating high-quality datasets that can adapt to diverse user profiles and preferences, thus enhancing the robustness of evaluations conducted on emerging task-oriented dialogue systems.

Noteworthy is the STAR dataset, designed to facilitate the transfer learning of models across different tasks and domains. STAR's schema-guided approach exploits a robust question-answering framework, thereby enabling dialogue systems to generalize more effectively when faced with novel tasks [55]. The flexibility offered by STAR demonstrates the shifts towards multi-domain adaptability, which is increasingly becoming a focal point in ongoing research efforts that leverage LLMs.

Further, advancements in creating contextually rich datasets have manifested in works such as the BiToD dataset, which introduces bilingual multi-domain dialogues, thus addressing the need for evaluating cross-lingual capabilities in task-oriented dialogue systems [56]. This dataset not only fosters performance metrics for bilingual systems but also encourages exploration in cross-domain transfer learning—an emerging challenge in the field.

In evaluating these datasets, the key trade-off remains between dataset generality and task specificity. While broad datasets such as MultiWOZ can offer insightful analyses into general dialogue system capabilities, they often miss subtleties that narrow-concept datasets encapsulate. Future research must strike a balance to develop hybrid datasets that maintain the depth required for nuanced evaluations while offering the breadth necessary for assessing adaptability and generalization capabilities.

As task-oriented dialogue systems continue to grow in complexity, addressing the emerging challenges of dataset creation is critical. Moving forward, researchers should prioritize the development of benchmarks that incorporate varied user intents, scenarios, and contexts while ensuring that the methodologies for dataset generation are scalable, reproducible, and capable of mitigating bias. By pursuing these avenues, the academic community can foster the development of more effective and equitable task-oriented dialogue systems, ultimately enriching the landscape of human-computer interactions.

4.4 Emerging Evaluation Approaches

Emerging evaluation approaches for task-oriented dialogue systems powered by large language models (LLMs) are gaining prominence in the ongoing quest to assess their performance effectively. Traditional evaluation methods, often reliant on isolated metrics such as accuracy or user satisfaction scores, inadequately capture the complex and dynamic nature of dialogic interactions. Consequently, innovative assessment frameworks that reflect the real-world nuances of conversational AI are becoming essential.

One significant trend is the increasing use of automated evaluation techniques that leverage LLMs for scoring dialogue quality. This paradigm shift recognizes the potential of LLMs to provide comparative assessments akin to human judgment. Recent studies have demonstrated that fine-tuning a dialogue evaluation model on dialogue data can yield reliable results that align closely with human annotations, thereby diminishing the need for extensive manual evaluations. Such methodologies could revolutionize the evaluation landscape by facilitating faster iterations in dialogue system development while ensuring quality through models like LLM-Eval, which aims to unify multiple dimensions of conversation quality in a single model call [57].

In addition, interactive evaluation frameworks are being increasingly utilized, representing an advancement over static, laboratory-based assessments. By simulating real-time conversational scenarios, these frameworks assess dialogue systems under conditions that closely mirror actual user interactions. Notably, the development of benchmarks such as MT-Eval emphasizes multi-turn interaction capabilities, enabling researchers to evaluate how well dialogue systems manage persistent context and user intent across extended conversations. This comprehensive approach reveals insights into model performance by addressing shortcomings observed in one-off turn evaluations, including error propagation and contextual relevance that significantly impact multi-turn dialogues [58].

Another noteworthy advancement is the emergence of meta-evaluation techniques that provide frameworks for assessing the efficacy of various evaluation metrics themselves. By analyzing the correlation between existing evaluation scores and human judgment, researchers can refine and mitigate biases inherent in traditional assessment methods, facilitating the establishment of more reliable evaluation standards. For instance, approaches that employ normalized metrics to adjust for differences in user expectations across demographic segments can yield a more equitable evaluation landscape, enhancing system fairness and accessibility [59].

Moreover, various strategies are being developed to augment dialogue systems with real-time data and external knowledge bases, necessitating a fresh evaluation perspective on how well these integrations enhance system responses. Research on Retrieval-Augmented Generation (RAG) highlights the importance of improving the robustness and reliability of responses while leveraging mechanisms that can provide contextually appropriate information from large databases [60]. Developing a comprehensive evaluation mechanism for such integrations remains critical, as the interplay between dialogue generation and information retrieval adds complexity to performance metrics.

As the field moves beyond traditional frameworks, there is a growing push towards establishing task-specific evaluation paradigms. These frameworks aim to tailor evaluation criteria to align with specific dialogue functionalities, acknowledging that a one-size-fits-all approach can distort insights. Task-oriented systems, whether applied in customer service, personal assistance, or casual dialogue contexts, require metrics that accurately reflect their unique interactions and user expectations. Consequently, a grounded understanding of user needs paired with contextual relevance metrics is shaping future evaluation methodologies [18].

Despite these advancements, emerging challenges persist. One pressing concern is the risk of evaluation metrics becoming overly entangled with specific system architectures, which could lead to potential overfitting and a loss of generalizability in assessments. Additionally, a lack of standardization in evolving benchmarks may create reproducibility issues across studies, complicating the ability to draw comparatives or aggregated insights. Therefore, establishing collaborative frameworks among researchers will be pivotal for standardizing evaluation metrics and ensuring a comprehensive understanding of LLMs in dialogue systems.

In conclusion, the advent of these innovative evaluation approaches is central to understanding the efficacy and user experience of task-oriented dialogue systems. The exploration of automated scoring, interactive assessments, and tailored evaluation methodologies are crucial for ongoing research aimed at refining LLM performance. By fostering robust evaluation techniques that resonate with real-world applications, the field can navigate the complexities inherent in dialogue systems and move towards more sophisticated and human-like interactions.

4.5 Challenges in Evaluation Metrics

Evaluating task-oriented dialogue systems presents a multitude of challenges that significantly impact their development and deployment. A critical challenge lies in defining appropriate evaluation metrics that comprehensively capture both the quantitative and qualitative aspects of user interactions. Conventional metrics like accuracy, precision, and recall are often insufficient, as they focus on individual turn correctness rather than the overall dialogue flow and success across multi-turn interactions. Such metrics may not reflect user satisfaction or the contextual coherence required for effective dialogue management, underscoring a fundamental disconnect between computational evaluations and user-centered experiences [61].

One primary concern is the presence of biases within training datasets, which invariably skew evaluation results and can lead to unfair system performance across diverse user demographics. The biases can originate from various sources, including the language styles prevalent in training materials and the contextual scenarios represented therein. These biases complicate the process of benchmarking dialogue systems against fairness and inclusiveness criteria, ultimately impeding efforts to build equitable AI solutions in dialogue [46]. Moreover, traditional evaluation methods often fail to account for cultural and contextual diversity in user interactions, which can significantly affect user perceptions and satisfaction.

Interpreting evaluation metrics themselves poses another challenge. High scores in traditional metrics might misleadingly suggest a system's proficiency in task completion, even when users report dissatisfaction or frustration due to irrelevant or inappropriate responses. For instance, an automated system excelling in achieving accuracy could still generate dialogues that users find unengaging or irrelevant [14]. This scenario underscores the necessity for nuanced metrics that integrate user feedback, such as user satisfaction scores or engagement metrics, into the evaluation process. Emerging trends advocate for the incorporation of user evaluations into automated metrics, allowing systems to better align with user expectations and engagement patterns [62].

Despite advancements in automated evaluation methods, including the use of large language models (LLMs) for dialogue quality scoring, considerable variability remains. For instance, different evaluators (human or model-based) can exhibit diverse performance when gauging instruction adherence, indicating the need for a more standardized, reliable assessment approach [63]. Automated evaluation frameworks that can adapt to the specifics of multi-turn dialogues would help address this variance and offer consistent assessments. Developing a unified evaluation paradigm that accommodates qualitative assessments alongside quantitative metrics remains pivotal for advancing the systematic evaluation of dialogue systems.

Furthermore, technical constraints, such as the scalability of evaluation methods, present practical barriers. As dialogue datasets grow in size and complexity, existing evaluation practices can become resource-intensive, slowing down the iterative improvement cycles essential for optimizing dialogue systems [64]. Addressing these scalability issues demands the development of efficient, automated metrics capable of handling large-scale data without compromising on depth or interpretability.

Looking ahead, researchers must also consider the evolving landscape of task-oriented dialogues shaped by advancements in LLMs. Future evaluation frameworks could benefit from a more interdisciplinary perspective, drawing from insights in user experience research and cognitive science to create systems capable of understanding and simulating human-like dialogue behaviors in varied contexts. By integrating principles from these fields, evaluation methodologies can introduce mechanisms to assess interaction richness, engage with user emotion, and measure contextual appropriateness. In doing so, the evaluation of task-oriented dialogue systems can evolve from a purely statistical exercise to a meaningful dialogue about user experience, engagement, and satisfaction—an essential shift for fostering user trust and promoting the sustainable deployment of AI in everyday applications [65].

In summary, overcoming these challenges mandates a concerted effort to rethink and redefine evaluation metrics for task-oriented dialogue systems. This effort should prioritize a balanced integration of quantitative and qualitative assessment measures, focus on minimizing bias, and develop scalable, user-centric evaluation methodologies that can adapt to the dynamic nature of dialogue interactions. A collaborative approach that leverages insights from various fields will be crucial in advancing evaluation strategies to ensure that task-oriented dialogue systems are robust, fair, and effective in meeting user needs.

5 Current Challenges in Implementation

5.1 Ethical Implications of Bias in Large Language Models

Bias in large language models (LLMs) poses significant ethical implications that profoundly affect user interactions and system efficacy in task-oriented dialogue systems. These biases often originate from the datasets used for training, which may reflect societal prejudices, stereotypes, or inequalities that inadvertently become embedded in model behaviors. As LLMs take on more roles in real-world applications, the consequences of these biases can manifest in harmful outputs that perpetuate discrimination and bias against marginalized communities.

The problem of bias in LLMs has been extensively documented in the literature. For instance, the research by [48] emphasizes that biases found in training datasets can lead to dialogues that are less favorable to certain demographic groups. Such biases can occur at multiple levels: from the representation of different identities within training data, to spoken or written language that reflects cultural stereotypes. The outcomes of these biases can have profound impacts on user interactions, shaping the experiences of individuals who engage with technical systems meant to assist rather than harm.

One critical consideration is how bias can lead to inequitable access to services provided by dialogue systems. An analysis of user interactions may reveal that certain groups receive less assistance or unsatisfactory responses to queries, which reflects a failure to adequately train the model on diverse inputs representative of all user demographics. The framework proposed by [66] illustrates these challenges, where the performance of dialogue systems can vary significantly depending on the user profile, potentially leading to systemic inequalities in service delivery.

Several approaches have been put forward to address these biases. Techniques such as data augmentation, where additional training data is generated or existing data is modified to reduce bias, and adversarial training methods, which involve training models against biased scenarios to improve robustness, have emerged as potential solutions. However, these strategies carry their own limitations. For example, while data augmentation can improve model exposure to diverse representations, it may not fully capture the complexities of real-world interactions. Furthermore, adversarial training may create models that can "sense" bias yet struggle to engage in genuinely fair behavior during practical deployment [3].

An emerging trend is the development of ethical datasets designed specifically to prevent bias in dialogue systems. As highlighted in [67], creating extensive, labeled datasets that account for diverse identities is crucial for building equitable AI systems. Implementing such robust datasets could mitigate biases in LLM outputs, fostering an environment where all users receive equitable treatment regardless of background.

Despite these advancements, the challenge persists in measuring and ensuring fairness effectively. Current approaches often rely on quantitative metrics that may overlook qualitative experiences of users. This raises important questions regarding interpretability and accountability in AI systems powered by LLMs [68]. Thus, a need arises for more transparent models that can articulate their decision-making processes, enhancing user trust and allowing for scrutinization of potential biases.

Ultimately, addressing bias in LLMs requires an interdisciplinary approach. Ethical considerations need to merge with technical fields, advocating for collaboration among AI researchers, ethicists, and sociologists. As we move forward, it becomes imperative for AI practitioners to not only focus on developing sophisticated models but also prioritize the ethical ramifications of their applications. Building fair and equitable systems must be at the forefront of research agendas to ensure fair user access and minimize harm in highly automated dialogue systems, shedding light on the interconnectedness between technology and social justice.

5.2 Interpretability and Transparency Challenges

Interpretability and transparency in large language models (LLMs) are critical factors influencing user trust and the overall efficacy of task-oriented dialogue systems. As these models increase in complexity, grasping their decision-making processes becomes increasingly difficult, presenting challenges for both developers and end-users. The opaque nature of LLMs leads to a reliance on "black boxes," leaving users unable to discern how input is transformed into output. This lack of clarity can significantly hinder effective model application, especially in sensitive areas such as healthcare or finance, where the stakes of decision-making are exceptionally high.

Efforts to enhance interpretability typically fall into three main categories: explainable AI (XAI) methods, model-agnostic techniques, and inherently interpretable model architectures. XAI techniques, such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations), analyze predictions by perturbing inputs and assessing changes in outputs. Through this approach, they provide insights into important features and contributing factors related to specific decisions [50]. However, these methods often rely on surrogate models, which can misrepresent the underlying complexities of LLMs, potentially leading to misleading interpretations of model behavior.

Model-agnostic approaches leverage visualization techniques to demystify LLM operations. For instance, attention visualizations provide insights into which tokens the model emphasizes when generating predictions. Yet, the interpretability of attention scores remains contentious, as they do not always correlate directly with a model's output quality or relevance [69]. Scholars such as Jain and Wallace have raised concerns about the clarity of attention weights, noting that high attention does not necessarily equate to importance, which can lead to overinterpretation of these metrics [70].

In pursuing inherently interpretable architectures, researchers are exploring alternatives such as decision trees or rule-based systems, which offer explicit and transparent decision processes. While these models facilitate easier understanding, they often fall short of the performance levels achieved by LLMs in complex tasks. Recent discussions advocate for a hybrid approach that retains the power of LLMs while integrating interpretable components, thereby fostering environments where explanations are naturally embedded into the modeling process [17].

Emerging trends emphasize the need to integrate transparency from the early design phases of LLMs. The development of systems featuring self-explainable components, where models can autonomously articulate their reasoning alongside predictions, has gained traction. Advancements in prompting techniques, such as Chain-of-Thought (CoT) prompting, demonstrate that structured reasoning prompts can yield outputs that reflect the logical steps in decision-making [71]. This suggests that user trust can be enhanced through mechanisms allowing LLMs to present their reasoning pathways, thereby improving user comprehension of automated interactions.

Nonetheless, challenges remain. Striking a balance between model performance and interpretability necessitates careful consideration; striving for transparency may inadvertently degrade performance or reintroduce bias issues lurking within the training data, undermining some anticipated benefits [72]. Additionally, user diversity means interpretability requirements can vary significantly—what provides clarity for one user may confuse another, complicating standardized interpretability metrics across applications and user demographics.

Looking forward, the AI research community must prioritize interpretability as an essential feature of models rather than a mere afterthought. Future research should focus on developing standardized frameworks for assessing interpretability, ensuring emerging dialogue systems meet ethical and practical transparency requirements. Engaging users continuously for feedback can also yield critical insights into the effectiveness of interpretability efforts. Through these initiatives, the dialogue systems of the future can embody not only advanced performance but also foster the trust and understanding necessary for empowering users in automated settings.

5.3 Resource Constraints and Accessibility Issues

The resource constraints and accessibility issues surrounding the deployment of large language models (LLMs) for task-oriented dialogue systems present significant challenges for organizations, particularly those with limited computational resources. The complexity and scale inherent in LLM architectures, such as GPT and BERT, entail substantial computational costs both for training and inference phases. These costs can hinder widespread adoption and exacerbate existing disparities in access to advanced AI technologies.

At the core of the resource challenge is the immense computational power required for training LLMs. Training these models necessitates not only advanced hardware—often involving multiple high-end GPUs or TPUs—but also substantial energy resources, which collectively contribute to high operational costs. For instance, the training of a model with billions of parameters can consume thousands of kilowatt-hours (KWh) of electricity, resulting in a notable carbon footprint, as discussed in [73]. As organizations increasingly adopt LLMs, the environmental impact associated with their training cycles has gained attention, prompting a push for more energy-efficient architectures and algorithms.

Accessibility issues arise from the combination of these computational demands with the financial barriers they impose. Smaller enterprises or non-profit organizations, which may lack the budgetary allocations necessary for extensive computational infrastructure, find it difficult to implement LLM-driven solutions. Many organizations also struggle with data scarcity, which hinders effective training and necessitates reliance on pre-trained models that can be prohibitively expensive to acquire or implement. Recent advances, such as task-optimized adapters and minimalist transfer learning methods, present promising avenues for reducing the computational burden, hence enhancing accessibility, but they often require expertise that smaller organizations may not possess [74].

Emerging trends in model compression and knowledge distillation have initiated promising alternatives to mitigate the resource intensity of deploying LLMs. Techniques such as these focus on retaining the performance of large models while vastly reducing their size and the computational resources required for inference [75]. For example, methods that utilize quantization and pruning can decrease the model size by significant margins while maintaining acceptable performance thresholds, allowing for deployment on consumer-grade hardware. However, the implementation of these methods still requires substantial initial development effort, which can discourage their adoption by organizations with limited research and development capabilities.

In light of these issues, innovative strategies are being explored to democratize access to LLMs. Open-source projects and collaborative frameworks that allow sharing of models and training protocols have emerged as effective ways to lower the barriers to entry for smaller organizations. The development of efficient inference toolkits, such as InfMoE, further underscores a commitment to addressing these challenges by enabling high-performance execution of LLMs on constrained hardware [76]. Moreover, initiatives that promote synthetic or augmented data generation, such as those leveraging LLMs as active annotators, can provide valuable resources in low-data regimes, thereby improving accessibility without the associated costs of traditional data collection [77].

In summary, the deployment of large language models in task-oriented dialogue systems is predominantly hindered by resource constraints and accessibility issues. While recent advancements aim to alleviate these burdens, substantial gaps remain, particularly for organizations with limited resources. As ongoing research focuses on efficiency and democratization of AI technologies, it will be critical to balance resource utilization with performance to ensure that the transformational potential of LLMs can be harnessed by a diverse array of users. Future research directions should strive to further refine model efficiency, explore novel methods of knowledge fusion, and leverage community-driven initiatives that can empower a broader range of societal stakeholders.

5.4 Balancing Automation with User Control

Balancing automation with user control in task-oriented dialogue systems presents a significant challenge, especially as these systems increasingly integrate large language models (LLMs) to enhance user interactions. While automated systems can efficiently handle numerous inquiries and tasks, excessive reliance on automation may compromise user satisfaction, engagement, and the ability to effectively address nuanced or complex inquiries. Therefore, it becomes crucial to explore various approaches to achieve a harmonious balance between automated assistance and user agency.

A pivotal strategy in this balance is the implementation of user feedback mechanisms. Systems should empower users to modify or redirect the dialogue flow as needed. For instance, allowing users to specify their preferences or provide real-time corrections greatly enhances the dialogue's effectiveness. Research has shown that incorporating feedback loops positively influences dialogue management; users feel more in control over automated interactions, leading to greater trust and satisfaction in the system [14]. This approach underscores the benefits of user control in enhancing task completion rates while maintaining user engagement.

In addition, adaptive control frameworks can significantly improve user interactions within task-oriented systems. These frameworks empower dialogue agents to dynamically adjust their response strategies based on user input and historical interaction data. For example, maintaining user profiles that capture preferences or previous interactions can inform how the system initiates dialogues or addresses user needs. This method provides automated responses while ensuring a degree of personalized interaction, showcasing both the efficiency of LLMs and the necessity of user-centric designs [23]. Indeed, while automation enhances efficiency, the ability to personalize interactions is fundamental to user retention and satisfaction.

However, it is essential to critically evaluate the limitations of automation. A pressing concern is the possibility of system overconfidence, where automated dialogues fail to accurately interpret user intent or context. Incorrect assumptions regarding a user's request can lead to frustration, particularly if the system lacks an easy mechanism for the user to regain control. Recent methodologies, such as Reinforcement Learning from Human Feedback (RLHF), have emerged to address these issues by allowing systems to learn from real-time user interactions and adapt their responses accordingly [24]. This reinforces the idea that user agency should be embedded at multiple levels within dialogue systems to ensure satisfactory interactions.

Emerging trends suggest that technologies like context-aware systems and knowledge graphs can further enhance this balance. By retaining information about previous interactions, these systems enable users to receive coherent continuations and contextually relevant responses without requiring extensive inputs beyond their initial queries. The integration of such mechanisms can foster a more seamless user experience, reducing frustration and facilitating deeper engagement in dialogues [22].

Moreover, the evolution of user control in automated systems must address ethical considerations. Clear guidelines and transparency protocols are vital to ensure users understand how their data and feedback are utilized. Additionally, mechanisms should be in place for users to provide feedback or report any errors or inappropriate responses generated by the system. Such measures help align automated systems with ethical standards and users' expectations, promoting a more trustworthy interaction landscape.

In conclusion, achieving a sustainable balance between automation and user control is essential for the ongoing development of task-oriented dialogue systems. The integration of adaptive learning mechanisms, effective feedback systems, and ethical guidelines will shape the future of these systems, ensuring that advancements in automation enrich rather than detract from user experiences. By developing systems that prioritize user control while leveraging the efficiencies provided by automation, developers can create enhanced and more satisfactory interaction paradigms for users, which ultimately lead to more robust and efficient dialogue systems. Future research should continue to investigate the intricacies of this balance, exploring how emerging technologies can better support user agency in diverse contexts.

5.5 Addressing Hallucinations and Miscommunication

Hallucinations in large language models (LLMs) refer to instances where the model generates output that includes factual inaccuracies, irrelevant information, or completely nonsensical statements, thereby impacting the reliability of dialogue systems. This phenomenon presents significant challenges in the context of task-oriented dialogue systems, where the accuracy and relevance of generated responses are crucial for effective user interaction.

The mechanisms underlying hallucinations can be traced to several factors, including the training data, model architecture, and extraction of contextual cues. The training process generally involves massive and diverse datasets, which, while rich in information, may also encompass inaccuracies and biases inherent in the source materials. Consequently, LLMs can produce errant outputs by overgeneralizing from these flawed examples. Research indicates that models like GPT-3 can exhibit pronounced overconfidence in their responses, often generating plausible but factually incorrect information simply because it fits the context syntactically rather than semantically. Moreover, models trained on datasets without stringent validation procedures are particularly susceptible to this issue, leading to what is referred to as "confidence without competence" [48].

Various strategies have been proposed to mitigate hallucinations, each with its strengths and limitations. One prominent approach is the integration of external knowledge sources to ground the model's output in verifiable information. For instance, augmenting LLMs with external databases, as explored in [4], has shown promise in enhancing the factual accuracy of generated dialogue. However, this reliance on external sources introduces additional complexities, such as the need for efficient retrieval mechanisms and real-time integration during dialogue, potentially affecting response times and system responsiveness.

Another technique involves refining the model's training methodology. For instance, fine-tuning with domain-specific data can enhance the model's contextual understanding, thus reducing hallucinations tied to specific domains. The work by [62] highlights the effectiveness of incorporating human feedback into the training loop, enabling the model to learn from its mistakes dynamically. However, the challenge remains in terms of scalability—balancing the contextual specificity with the model's overall generalization capabilities across diverse tasks continues to prompt research [78].

Model interpretability also plays a critical role in addressing hallucinations. Techniques such as explainable AI (XAI) can foster transparency in decision-making processes of LLMs, allowing developers and users alike to understand the rationale behind certain outputs. Understanding these mechanisms is key for improving user trust and enhancing dialogue quality [61]. Conversely, the complexity of these models often obscures their internal workings, rendering the identification and correction of errors a challenging proposition.

Emerging trends focus on preemptively addressing hallucinations through proactive dialogue management strategies. Approaches that involve a continuous feedback loop where user interactions are analyzed to inform subsequent model updates seem promising. Such frameworks can allow for the adaptation of dialogue systems that learn from live interactions, thus gradually minimizing the frequency of hallucinations [79].

In conclusion, as research continues to elucidate the intricacies surrounding hallucinations in LLMs, the imperative remains for a multifaceted approach that combines the refinement of training methodologies, the integration of external knowledge bases, and enhanced interpretability mechanisms. This synthesis not only addresses the immediate concerns of hallucinations but also lays the foundation for more robust and reliable task-oriented dialogue systems, propelling the technology toward greater effectiveness in real-world applications. Seeking effective solutions will require ongoing collaboration across disciplines, drawing insights from user experience, AI ethics, and cognitive science to create models that are both powerful and responsible.

5.6 Future Challenges in Regulation and Compliance

The deployment of large language models (LLMs) in task-oriented dialogue systems introduces significant regulatory and compliance challenges. As these models increasingly dominate interaction across various industries, the need for robust frameworks governing their use becomes paramount. The integration of sophisticated AI systems with sensitive user data necessitates mechanisms ensuring privacy, accountability, and adherence to evolving regulatory expectations.

Central to these regulatory challenges is the requirement to comply with data privacy laws such as the General Data Protection Regulation (GDPR) in Europe and the California Consumer Privacy Act (CCPA) in the United States. These regulations mandate explicit user consent regarding the processing of personal data. However, the opaque nature of LLMs raises questions about accountability for entities deploying these systems. Notably, concerns emerge around the storage and utilization of dialogue data without clear user consent, potentially resulting in violations of data protection regulations. To mitigate this risk, implementing "privacy by design" in LLM frameworks—ensuring that user privacy considerations are integrated from the outset rather than treated as an afterthought—is essential [15].

Another critical factor is the challenge of understanding and documenting the decision-making pathways of LLMs. The complex algorithms that underpin these models can lead to unpredictable outcomes, complicating compliance with regulations that demand transparency. Initiatives geared toward explainable AI (XAI) are making strides in addressing these issues by developing models capable of articulating their reasoning. Such advancements not only bolster user trust but also facilitate compliance with regulatory frameworks that require AI systems to be explainable. For instance, the incorporation of guiding frameworks like decision trees or attention visualization can illuminate how LLMs arrive at specific responses during user interactions, aligning with the transparency standards set forth by various regulations [80].

Moreover, compliance is inextricably linked with the ethical implications of utilizing LLMs in dialogue systems. These models have historically been trained on vast datasets, often containing biased or unfiltered information. Such biases can inadvertently lead to discriminatory practices through dialogue systems. Consequently, developers and organizations must actively work to identify, mitigate, and disclose biases that emerge from their models. Implementing bias-detection algorithms and conducting regular audits of outputs for fairness and inclusivity are vital measures for ensuring compliance with ethical standards, resonating with values embedded in contemporary regulatory frameworks [81].

Emerging trends in regulation highlight an increasing focus on the ethical deployment of AI technologies, reflecting a broader societal push for accountability and moral responsibility in machine learning. Governments and regulatory bodies are progressively seeking collaboration with AI researchers, industry leaders, and ethicists to establish comprehensive guidelines for the use of LLMs in real-world applications. Such collaboration will be crucial for developers to create adaptable regulatory frameworks capable of evolving alongside the rapid advancement of technology and its integration into societal functions.

As organizations navigate these challenges, a fundamental strategy lies in fostering a culture of compliance that emphasizes continuous education regarding regulatory requirements and ethical AI practices. Integrating training programs that inform employees about the implications of data privacy laws and ethical AI can promote informed decision-making and responsible model deployment. Research indicates that organizations that prioritize compliance education tend to achieve better outcomes while reaping the benefits of AI technologies [46].

In conclusion, the future regulatory landscape for large language models within task-oriented dialogue systems is likely to be marked by increased scrutiny and a growing demand for compliance with rigorous data protection standards and ethical considerations. By adopting proactive measures, fostering collaboration, and enhancing transparency, organizations can effectively navigate the complex interplay of technology and regulation. As these trends unfold, the ongoing dialogue surrounding regulation and compliance will profoundly influence the development trajectory of LLMs, shaping their adoption and integration into society.

6 Future Directions and Emerging Trends

6.1 Multimodal Integration in Task-Oriented Dialogue Systems

The integration of multimodal capabilities in task-oriented dialogue systems represents a promising direction for enhancing user interaction through the simultaneous processing of diverse input forms, including text, speech, and images. This approach aims to create more intuitive and contextually aware dialogue agents that can adapt to varying user needs and environmental inputs. Given the rapid advancements in large language models (LLMs), such as those seen in recent studies, the incorporation of multiple modalities into dialogue systems aligns well with the growing expectation for conversational agents to perform complex, contextually rich operations.

Multimodal dialogue systems leverage the strengths of different input types, for example, by combining textual commands with visual inputs to create a more holistic understanding of user intent. The ability to parse and integrate data from multiple sources fosters an enriched interaction experience, potentially leading to higher user satisfaction and improved task completion rates. Consider the work of recent frameworks instilling visual grounding in dialogue interactions, where agents can utilize images alongside textual queries to enhance their responses, reflecting a more clear understanding of the context or scenario presented by the user [6]. In these systems, the interplay between visual and textual data can be modeled using attention mechanisms that allow the dialogue agent to focus on relevant details from either modality as required.

A crucial element of multimodal integration lies in the representation of different data types to facilitate seamless interaction. Representation learning plays a vital role in ensuring that various inputs—whether from dialogue turns, visual data from images, or audio cues—are converted into a unified feature space where they can be effectively processed. Techniques like joint embedding spaces aid in this endeavor, enabling the model to learn coherent representations that can bridge the gap between modalities, as highlighted by various empirical studies [18]. Furthermore, specialized architectures that utilize transformer-based structures have shown promise in managing multimodal interactions by facilitating the integration of context from different inputs through comprehensive attention mechanisms.

However, despite these advancements, challenges persist, particularly in the complexity of model training and computational efficiency. The need for extensive datasets that embody multimodal interactions is crucial for training robust dialogue systems. While existing datasets such as MultiWOZ primarily focus on dialogue without incorporating multimodal elements, initiatives to develop richer datasets that include visual and audio inputs are emerging as fundamental to progress in this domain [67]. Moreover, integrating these varied data types incurs additional computational costs and necessitates sophisticated model architectures that can handle real-time processing, posing scalability issues in practical deployments.

Emerging trends indicate a shift towards incorporating real-world scenarios into dialogue system training, allowing for improved handling of noise and variability inherent in user interactions across modalities. Techniques such as user simulation, drawing from LLMs to mimic diverse interaction styles, promise to enrich the training landscape by generating realistic multimodal dialogues [66]. Furthermore, the use of reinforcement learning methodologies to optimize dialogue flow in response to multimodal cues showcases the potential for dynamic adaptability, enhancing system robustness [15].

In summation, the integration of multimodal capabilities in task-oriented dialogue systems holds significant potential for advancing user interaction quality and broadening the operational scope of conversational agents. Continued research into effective representation learning, dataset development, and model architectures will be pivotal in overcoming existing limitations, enabling systems to understand and engage users through diverse and complex modalities. The trajectory indicates that future innovations will not only focus on enhancing interactivity but will also explore the ethical implications of such systems to ensure equitable user experiences across varied contexts.

6.2 Adaptive Learning Mechanisms

Adaptive learning mechanisms play a pivotal role in enhancing the responsiveness and personalization of task-oriented dialogue systems, effectively allowing these systems to refine their interactions based on continuous user feedback and evolving requirements. This adaptability contributes significantly to user satisfaction by making dialogue agents more attuned to individual preferences and interaction histories, thereby fostering a more engaging conversational experience.

One promising approach to adaptive learning in dialogue systems is reinforcement learning (RL), particularly when combined with user feedback. By applying RL algorithms, dialogue agents can receive rewards for successful interactions, optimizing their responses based on real-world outcomes. For instance, deep reinforcement learning techniques enable models to evaluate potential future rewards across multiple dialogue turns, thereby reducing the shortsightedness typical of traditional models that optimize only for immediate response generation. Such methods have shown success in improving coherence and user engagement in conversations, as demonstrated by studies integrating RL techniques with dialogue generation systems [8].

Moreover, the integration of personalization techniques, such as memory-augmented learning, allows dialogue systems to retain relevant information about users' preferences and past interactions. This is particularly effective in scenarios where the user-agent relationship is long-term, enabling the system to adapt its responses according to the historical context. For example, the Memory-Augmented Dialogue Management model combines local understanding of the current user utterance with global context from the entire dialogue history, thereby enhancing the relevance of generated responses [23]. Such architectures reveal the profound impact of previous interactions on user expectations and satisfaction, informing the tailoring of subsequent responses.

Emerging adaptive models further explore lifelong learning paradigms, enhancing systems' capacity to continuously learn from user interactions over extended periods. Lifelong learning frameworks help avoid catastrophic forgetting, where performance on older tasks diminishes as the model adapts to new user behavior. Techniques such as elastic weight consolidation or progressive neural networks may be employed to strike a balance between improving performance on new tasks while preserving essential knowledge derived from past interactions [53]. These strategies facilitate sophisticated user engagement by allowing the model to retain critical data while integrating new information.

Another innovative approach involves dynamic adaptations based on situational context. Contextual adaptation can be managed using real-time analytics to assess user sentiment and modify dialogues accordingly. This mechanism not only enriches the conversational experience but also enables the system to adjust its behavior dynamically based on user satisfaction metrics and emotional cues, enhancing overall user interaction [82]. The ability to interpret user intent and detect emotional states through natural language input can significantly propel the development of more intuitive dialogue systems.

Despite these advancements, challenges persist. The trade-offs between personalization and computational efficiency pose significant hurdles in large-scale deployment. Models engaged in continual learning often encounter increased computational demands and the risk of information overload, where excessive retained information may lead to confusion in interactions. Additionally, ensuring the privacy and security of user data in personalized systems remains a crucial ethical consideration, necessitating ongoing research to develop frameworks that facilitate effective and secure user experience customization.

In conclusion, as adaptive learning mechanisms continue to foster environments conducive to personalization within dialogue systems, further exploration into methods that enhance computational efficiency, safeguard user privacy, and adopt effective lifelong learning strategies will be essential. The interplay of reinforcement learning, memory utilization, and real-time contextual adaptation signifies a promising trajectory for the evolution of task-oriented dialogue systems, ultimately fostering richer and more meaningful interactions between users and their digital assistants [18], [83].

6.3 Interdisciplinary Approaches to Dialogue Design

Interdisciplinary collaboration presents a significant frontier for advancing task-oriented dialogue systems by weaving together insights from cognitive science, linguistics, and artificial intelligence. These domains offer unique perspectives that enhance the design and efficacy of dialogue systems, fostering interactions that are not only technically robust but also aligned with human communicative patterns and cognitive behavior.

Cognitive science provides valuable frameworks for understanding how humans process information during dialogues. By integrating cognitive models, dialogue systems can leverage principles from human reasoning and decision-making, enhancing their ability to predict user intents and adapt responses accordingly. For instance, research indicates that employing cognitive architectures can improve user-agent interaction efficiency by replicating human-like reasoning processes in dialogue management systems [55]. This approach facilitates naturalistic interactions and reduces user frustration, a common issue in conventional models that fail to interpret nuanced user queries.

Linguistics contributes fundamentally to the natural language processing (NLP) component of dialogue systems, influencing how systems understand and generate language. Insights into syntax, morphology, and semantics enable the development of models that appreciate the subtleties of human language, such as idiomatic expressions and contextual relevance, which may significantly differ across dialects and cultural backgrounds [54]. The application of linguistic theories allows dialogue systems to produce contextually accurate and culturally sensitive responses. Moreover, with knowledge-grounded dialogue generation, systems can incorporate domain-specific language that resonates better with users [84].

Furthermore, interdisciplinary approaches also encourage the coalescence of machine learning techniques with cognitive and linguistic insights to develop more sophisticated learning paradigms for dialogue systems. For instance, integrating reinforcement learning with cognitive models allows systems to adapt based on real-time user feedback while considering cognitive biases in user decision-making processes. Techniques such as transfer learning from related cognitive tasks to enable better dialogue responses under low-data scenarios exemplify this synergy [19]. While these methods show promise, they also highlight challenges, such as effectively mapping the complexities of human cognition to computational models, which can lead to oversimplification if not approached with precision and care.

Emerging trends showcase the increasing relevance of personalization within task-oriented dialogue systems, driven by insights from both cognitive psychology and user experience (UX) design. Personalizing dialogue agents to align with individual user preferences has become a focal research area, with models leveraging past interactions to provide tailored responses. However, as illustrated in recent studies, this also introduces the risk of overfitting to individual user data, challenging the balance between personalized and generalizable solutions [3].

Another concern is the trade-off between system complexity and interpretability. More sophisticated models that utilize cognitive and linguistic insights are inherently complex, which may hinder users' trust and understanding of system decisions. To mitigate this, implementing explainable AI principles that demystify decision processes is crucial. Developing interfaces that communicate the reasoning behind specific responses can bridge the interpretability gap, as suggested by recent advancements in dialog policy learning [20].

In conclusion, fostering interdisciplinary approaches in dialogue design enhances the potential to create more effective, user-friendly, and context-aware systems. By synthesizing insights from cognitive science and linguistics with advancements in AI, researchers can address both the technical challenges and human-centric aspects of dialog systems, paving the way for future innovations. As the field progresses, continued exploration of these interdisciplinary practices will be essential for developing next-generation dialogue systems that can engage users in more meaningful and effective ways.

6.4 Efficient Evaluation Frameworks

Efficient evaluation frameworks are critical for assessing the performance of task-oriented dialogue systems powered by large language models (LLMs). As these systems increasingly incorporate sophisticated dialogue strategies, it becomes essential to analyze both quantitative and qualitative aspects of model outputs to ensure they meet user expectations and operational benchmarks. Traditional evaluation methods have often emphasized quantitative performance metrics, such as accuracy, F1-score, and dialogue success rates, which evaluate specific capabilities of a dialogue system in isolation. However, such metrics may fail to capture the intricacies of user experience, particularly in multi-turn dialogues where the context significantly influences dialogue flow and coherence. Thus, emerging evaluation frameworks seek to present a more holistic view by incorporating qualitative assessments that reflect the nuances of human-computer interaction.

Recent studies have highlighted the limitations of standard metrics like accuracy, which can overlook contextual relevance and emotional engagement in user responses. To address this gap, the incorporation of user feedback and emotional valence metrics provides richer insights into user satisfaction and dialogue effectiveness, rather than solely focusing on task completion rates [85]. Moreover, innovative frameworks such as MT-Eval have been developed to assess multi-turn interactions by categorizing dialogues into distinct patterns like recollection, expansion, and refinement, thereby offering a multidimensional perspective on conversational capabilities [58]. This evolution from traditional single-turn evaluations to comprehensive multi-turn assessments underscores the need for adaptable and nuanced evaluation strategies in dialogue systems powered by LLMs.

A key aspect of developing efficient evaluation frameworks lies in the integration of user-centric metrics, which include satisfaction scores, retention rates, and perceived helpfulness of responses. These metrics align dialogue performance with actual user needs while providing actionable feedback for model improvement. Particularly, LLMs, with their ability to generate context-aware responses, can be evaluated using mixed methodologies that leverage both qualitative user study insights and quantitative model performance metrics [86]. Advanced approaches incorporating automated evaluation techniques powered by LLMs themselves show promise in enhancing efficiency in this domain. For instance, leveraging LLMs to perform dialogue assessments that closely mirror human judgment allows for rapid iterations and refinements in model training, circumventing the exhaustive process of manual evaluation.

As dialogue systems evolve, so too do the evaluation challenges that researchers face. Current benchmarks often struggle to encapsulate the dynamic nature of conversational interactions influenced not only by model outputs but also by varying user intents and contextual shifts. In response to this, techniques such as retrieval-augmented evaluation frameworks, where contextual embeddings improve understanding and relevance, have emerged from recent research and demonstrate potential in bridging these gaps [87]. Simultaneously, the importance of developing benchmarks that can simulate real-world scenarios is emphasized, with frameworks like Loong providing structured approaches for evaluating long-context understanding in dialogues [88].

Furthermore, the advancement of models capable of self-adaptation and memory retention suggests that future evaluation approaches must also focus on metrics that assess long-term interactions and memory utilization effectively. Understanding how dialogue agents learn and adapt over time is critical for establishing effectiveness in protracted user engagement scenarios. For example, frameworks incorporating multi-session conversational testing and real-time feedback loops may provide insights not only into static performance but also into adaptability and contextual relevance across multiple interactions.

In conclusion, fostering a standardized approach to evaluation methodologies that encompasses both quantitative and qualitative dimensions will significantly enhance the robustness and applicability of task-oriented dialogue systems. Future research should prioritize the integration of context-aware assessment frameworks, the development of efficient automated evaluation mechanisms, and the creation of benchmarks that capture the full dynamism of human conversation. Addressing these areas will not only improve system performance but also facilitate deeper user engagement, ultimately leading to more sophisticated and human-like AI interactions in dialogue systems.

6.5 Challenges with Ethical Considerations and Bias

The deployment of large language models (LLMs) in task-oriented dialogue systems raises significant ethical challenges and concerns regarding bias. As these systems are integrated into increasingly sensitive applications, the implications of their outputs become critical. Inherent biases embedded within training data can lead to skewed representations and reinforce societal stereotypes, impacting fairness and equity in user interactions. For instance, findings suggest that LLMs trained on uncurated datasets may perpetuate negative stereotypes against specific demographic groups, leading to discriminatory dialogue outputs [48].

The root of these biases often lies in the training datasets themselves, which may reflect historical imbalances present in society. As LLMs learn from vast corpora that capture language usage patterns, any biases found within these data can be assimilated into model behaviors [89]. A recent examination highlights how LLMs may inadvertently engage in biased dialogue, particularly in scenarios requiring empathy or nuanced understanding, resulting in responses that can alienate marginalized voices [4].

Currently, various approaches have been proposed to mitigate these biases. One promising avenue involves augmenting training datasets with diverse and ethically sourced examples, which serve to balance representation across different demographic groups. However, this method presents challenges; careful curation is necessary to ensure that augmented data retains its integrity and does not inadvertently introduce new biases [61]. Moreover, training approaches such as reinforcement learning from human feedback (RLHF) can be employed to align LLM outputs with user expectations, yet they necessitate substantial amounts of reliable annotated data, which may not be available for all domains [44].

Addressing transparency in LLM dialogue generation is also intertwined with ethical considerations. As these models are often perceived as 'black boxes', their decision-making processes lack interpretability, making it difficult for users to understand the basis of generated dialogues. The challenge of transparency is paramount, especially when biases could lead to harmful or untrustworthy interactions [90]. Enhanced interpretability can be pursued through techniques from explainable AI (XAI), allowing models to provide rationales for their decisions, thereby fostering user trust and understanding [91].

Emerging trends in the field indicate a growing focus on frameworks that promote ethical AI while balancing performance and accountability. For instance, various research initiatives are advocating for the establishment of norms and standards that guide the ethical deployment of LLMs in sensitive domains, which may include legal, healthcare, or children's education [61]. Additionally, recent studies have proposed the incorporation of ethical auditing mechanisms into LLM training and evaluation processes, emphasizing the importance of regular assessments to identify and rectify biases [46].

Looking forward, the integration of multimodal approaches presents an innovative perspective for enhancing the ethical frameworks of dialogue systems. By combining textual inputs with visual or auditory cues, dialogue models could potentially achieve a deeper contextual understanding, thereby mitigating miscommunication risks and enhancing the accuracy of user intent recognition [27]. Furthermore, the adaptation of continual learning paradigms within LLMs may allow for dynamic updating processes that respond to shifting societal values and expectations, thereby reducing the potential for outdated or biased outputs over time [92].

In conclusion, the ethical considerations and biases inherent in deploying large language models for task-oriented dialogue systems necessitate rigorous attention from researchers and practitioners alike. A balanced approach that harmonizes model performance with ethical obligations, transparency, and inclusivity can pave the way for more equitable and effective interactive systems. As the field progresses, prioritizing ethical frameworks alongside technical advancements will be crucial in building dialogue systems that genuinely serve diverse communities.

6.6 Emerging Technologies for Enhanced Interaction

The integration of emerging technologies into task-oriented dialogue systems holds the promise of transforming user interactions and significantly enhancing system capabilities. Recent advancements in natural language processing, machine learning, and related fields have opened new avenues for improving dialogue systems' adaptability, contextual understanding, and overall performance. Key emerging technologies in this domain include retrieval-augmented generation (RAG), reinforcement learning from human feedback (RLHF), and memory-augmented architectures, each contributing uniquely to the evolution of interactive experiences.

Retrieval-augmented generation is increasingly recognized as a transformative innovation in mitigating the limitations of large language models (LLMs). By enriching the generative capabilities of these models with external knowledge retrieval mechanisms, RAG can provide users with up-to-date and contextually relevant information, thereby enhancing the accuracy and informativeness of responses. Recent investigations highlight how RAG systems effectively reduce instances of hallucination in LLMs by grounding outputs in factual data retrieved from maintained databases [4]. Furthermore, the flexibility introduced by partitioned memory structures in RAG systems allows specific memories to be prioritized over others, facilitating targeted interactions that boost user satisfaction [93]. However, challenges remain in optimizing retrieval strategies without compromising the coherence of generated text, necessitating ongoing research into the balance between retrieval fidelity and generation fluency.

Reinforcement learning techniques, particularly those informed by human feedback, represent another promising avenue for continuously enhancing dialogue agents. These techniques enable models to align more closely with user expectations and preferences, resulting in more personalized and engaging interactions. Studies reveal that dialogue agents employing RLHF strategies can better grasp context and user intent during extended interactions, thereby improving conversational quality [16]. Nevertheless, the implementation of RLHF presents challenges, especially concerning effective incorporation of human evaluative feedback into training cycles. Recent innovations focus on developing robust frameworks that allow models to learn from user interactions in a structured way, thus enhancing their adaptability in dynamic conversational environments [18].

Memory-augmented architectures signify another critical frontier for enriching task-oriented dialogues. These systems incorporate explicit memory components that enable dialogue agents to retain and utilize information across extensive conversation histories. This capability enhances an agent's contextual awareness, empowering systems to deliver coherent responses based on accumulated user data. For instance, memory architectures that distinguish between short-term and long-term memory can significantly elevate an agent's conversational consistency and relevance, enabling interactions that closely resemble human communication [21]. However, designing efficient memory management protocols poses a notable challenge, particularly in harmonizing the model's computational efficiency with the depth of contextual understanding.

Moreover, technologies that incorporate multimodal inputs are gaining traction, facilitating the development of more interactive and engaging dialogue systems. By merging text with audio and visual data, these systems enhance their ability to interpret user intents and provide richer responses. The relevance of visual context in dialogue interacts effectively with user engagement, demonstrating that multimodal approaches can bridge the gap between human and machine interaction more seamlessly [94]. Nevertheless, the complexity of managing and synchronizing diverse input types remains a technical barrier that requires innovative solutions.

The collective momentum generated by these emerging technologies indicates a trend toward more sophisticated and user-centric dialogue systems. As these technologies continue to evolve, their integration will necessitate a deeper understanding of the interactive dynamics present within dialogue systems. Future research directions could focus on refining these systems to effectively handle ambiguous user queries, employing advanced contextual embeddings to enhance understanding across dialogue turns, and developing standardized evaluation metrics to measure the effectiveness of these integrated approaches. By systematically addressing the inherent challenges posed by these technologies, researchers can pave the way for significantly enhanced task-oriented dialogue systems that deliver meaningful and efficient user interactions.

7 Conclusion

The exploration of large language models (LLMs) within task-oriented dialogue systems has evolved significantly, exhibiting transformative capabilities alongside persistent challenges. Central to the realization of robust dialogue systems is the ability of LLMs to generate contextually aware and coherent responses, effectively addressing user intents across diverse applications. Recent advancements, exemplified by models such as GPT and its subsequent iterations, have shown that LLMs are capable of learning from vast datasets, thus enabling them to excel in various dialogue-related tasks. For instance, methodologies like reinforcement learning from human feedback have enhanced the responsiveness and user alignment of these systems, illustrating the effectiveness of human-centered training approaches [44].

While the progress in utilizing LLMs has been substantial, challenges remain. One significant concern is the tendency of LLMs to generate responses that may lack factual accuracy or exhibit bias, issues that have implications for fairness and reliability in dialogue interactions [95]. The interaction of these models with external knowledge bases offers a pathway to mitigate such limitations, as highlighted by the integration of retrieval-augmented generation approaches [4]. These developments showcase the continuous need for research focused on both improving the foundational capabilities of LLMs and addressing their inherent biases.

Comparatively, traditional task-oriented dialogue systems often struggle with modularization limitations and error propagation, which can degrade overall system performance [47]. End-to-end models have emerged as a solution to these challenges by streamlining the learning process and allowing for more coherent interactions. The shift towards hybrid approaches that combine the strengths of modular architectures with end-to-end learning paradigms is gaining momentum, as demonstrated in studies advocating for less rigid dialogue management structures [62]. This hybridization represents an emerging trend that aims to leverage the best features of both approaches, facilitating greater flexibility and adaptability in dialogue systems.

Emerging trends indicate a growing interest in multimodal integration, where text-based dialogue systems are enhanced with audiovisual inputs, thereby diversifying user interaction modalities and enriching conversational experiences. This integrative approach can foster deeper contextual understanding and engagement, echoing findings from research emphasizing the role of context-aware systems [94]. Moreover, the increasing importance of personalization in dialogue systems highlights the necessity for adaptive learning mechanisms that can modify system responses based on user history [66]. This trend presents both an opportunity and a challenge, as organizations must consider how to efficiently implement personalization while managing user privacy concerns.

As the field progresses, future research must prioritize the exploration of ethical implications and develop frameworks for transparent and interpretable AI systems [5]. The need for systematic evaluation methodologies cannot be overstated, particularly in light of the pressing challenges associated with assessing LLM outputs across various tasks [96]. Establishing rigorous benchmarks and evaluation criteria will be vital in advancing the capabilities of dialogue systems while ensuring accountability.

In conclusion, the landscape of task-oriented dialogue systems powered by LLMs presents a dynamic interplay of innovation and challenge. While the advancements are promising, continued exploration will be essential to unlock the full potential of these models. This includes addressing biases, enhancing factual accuracy, and developing comprehensive evaluation strategies that adhere to high ethical standards. Future research directions should focus on integrating diverse methodologies and fostering interdisciplinary collaboration, ensuring that the evolution of dialogue systems remains aligned with user needs and societal values. The path forward holds significant promise for advancing our understanding and implementation of intelligent conversational agents, with the potential to transform user interactions across various domains.

REFERENCES

[1] J. Wei, X. Wang, D. Schuurmans, M. Bosma, E. Chi, F. Xia, Q. Le, and D. Zhou, "Chain of thought prompting elicits reasoning in large language models," *ArXiv*, vol. abs/2201.11903, 2022. 1, 2

[2] Z. Hu, Y. Feng, A. Luu, B. Hooi, and A. Lipani, "Unlocking the potential of user feedback: Leveraging large language model as user simulators to enhance dialogue system," *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 2023. 1

[3] M. Henderson, I. Vulic, D. Gerz, I. Casanueva, P. Budzianowski, S. Coope, G. P. Spithourakis, T.-H. Wen, N. Mrksic, and P. hao Su, "Training neural response selection for task-oriented dialogue systems," *ArXiv*, vol. abs/1906.01543, 2019. 1, 3, 11, 13, 19

[4] B. Peng, M. Galley, P. He, H. Cheng, Y. Xie, Y. Hu, Q. Huang, L. Lidén, Z. Yu, W. Chen, and J. Gao, "Check your facts and try again: Improving large language models with external knowledge and automated feedback," *ArXiv*, vol. abs/2302.12813, 2023. 1, 2, 5, 8, 9, 10, 16, 20, 21

[5] J. Kaddour, J. Harris, M. Mozes, H. Bradley, R. Raileanu, and R. McHardy, "Challenges and applications of large language models," *ArXiv*, vol. abs/2307.10169, 2023. 1, 21

[6] R. Shu, E. Mansimov, T. Alkhouli, N. Pappas, S. Romeo, A. Gupta, S. Mansour, Y. Zhang, and D. Roth, "Dialog2api: Task-oriented dialogue with api description and example programs," *ArXiv*, vol. abs/2212.09946, 2022. 1, 17

[7] I. Serban, A. Sordoni, Y. Bengio, A. C. Courville, and J. Pineau, "Building end-to-end dialogue systems using generative hierarchical neural network models," in *AAAI Conference on Artificial Intelligence*, 2015, pp. 3776–3784. 2, 10

[8] J. Li, W. Monroe, A. Ritter, D. Jurafsky, M. Galley, and J. Gao, "Deep reinforcement learning for dialogue generation," *ArXiv*, vol. abs/1606.01541, 2016. 2, 18

[9] K. Sun, S. Moon, P. A. Crook, S. Roller, B. Silvert, B. Liu, Z. Wang, H. Liu, E. Cho, and C. Cardie, "Adding chit-chat to enhance task-oriented dialogues," in *North American Chapter of the Association for Computational Linguistics*, 2020, pp. 1570–1583. 2

[10] Y. Su, L. Shu, E. Mansimov, A. Gupta, D. Cai, Y.-A. Lai, and Y. Zhang, "Multi-task pre-training for plug-and-play task-oriented dialogue system," in *Annual Meeting of the Association for Computational Linguistics*, 2021, pp. 4661–4676. 2, 10

[11] S. Yuan, H. Zhao, S. Zhao, J. Leng, Y. Liang, X. Wang, J. Yu, X. Lv, Z. Shao, J. He, Y. Lin, X. Han, Z. Liu, N. Ding, Y. Rao, Y. Gao, L. Zhang, M. Ding, C. Fang, Y. Wang, M. Long, J. Zhang, Y. Dong, T. Pang, P. Cui, L. Huang, Z. Liang, H. Shen, H. Zhang, Q. Zhang, Q. Dong, Z. Tan, M. Wang, S. Wang, L. Zhou, H. Li, J. Bao, Y. Pan, W. Zhang, Z. Yu, R. Yan, C. Shi, M. Xu, Z. Zhang, G. Wang, X. Pan, M.-J. Li, X. Chu, Z. Yao, F. Zhu, S. Cao, W. Xue, Z. Ma, Z. Zhang, S. Hu, Y. Qin, C. Xiao, Z. Zeng, G. Cui, W. Chen, W. Zhao, Y. Yao, P. Li, W. Zheng, W. Zhao, Z. Wang, B. Zhang, N. Fei, A. Hu, Z. Ling, H. Li, B. Cao, X. Han, W. Zhan, B. Chang, H. Sun, J. Deng, J. Li, L. Hou, X.-M. Cao, J. Zhai, Z. Liu, M. Sun, J. Lu, Z.-G. Lu, Q. Jin, R. Song, J. Wen, Z. Lin, L. Wang, H. Su, J. Zhu, Z. Sui, J. Zhang, Y. Liu, X. He, M. Huang, J. Tang, and J. Tang, "A roadmap for big model," *ArXiv*, vol. abs/2203.14101, 2022. 2

[12] Y. Qian, W. Zhang, and T. Liu, "Harnessing the power of large language models for empathetic response generation: Empirical investigations and improvements," in *Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 6516–6528. 2

[13] A. Mosharrof, M. H. Maqbool, and A. Siddique, "Zero-shot generalizable end-to-end task-oriented dialog system using context summarization and domain schema," *ArXiv*, vol. abs/2303.16252, 2023. 2

[14] J. Williams and G. Zweig, "End-to-end lstm-based dialog control optimized with supervised and reinforcement learning," *ArXiv*, vol. abs/1606.01269, 2016. 3, 4, 6, 8, 12, 15

[15] N. Bang, J. Lee, and M. Koo, "Task-optimized adapters for an end-to-end task-oriented dialogue system," *ArXiv*, vol. abs/2305.02468, 2023. 3, 7, 9, 16, 17

[16] D. Ulmer, E. Mansimov, K. Lin, J. Sun, X. Gao, and Y. Zhang, "Bootstrapping llm-based task-oriented dialogue agents via self-talk," *ArXiv*, vol. abs/2401.05033, 2024. 3, 6, 21

[17] V. Hudecek and O. Dusek, "Are llms all you need for task-oriented dialogue?" *ArXiv*, vol. abs/2304.06556, 2023. 3, 14

[18] Z. Yi, J. Ouyang, Y. Liu, T. Liao, Z. Xu, and Y. Shen, "A survey on recent advances in llm-based multi-turn dialogue systems," *ArXiv*, vol. abs/2402.18013, 2024. 3, 5, 12, 17, 18, 21

[19] K. Mo, Y. Zhang, S. Li, J. Li, and Q. Yang, "Personalizing a dialogue system with transfer reinforcement learning," in *AAAI Conference on Artificial Intelligence*, 2016, pp. 5317–5324. 3, 11, 19

[20] B. Liu, G. Tür, D. Z. Hakkani-Tür, P. Shah, and L. Heck, "End-to-end optimization of task-oriented dialogue model with deep reinforcement learning," *ArXiv*, vol. abs/1711.10712, 2017. 3, 19

[21] R. R. G. Reddy, D. Contractor, D. Raghu, and S. Joshi, "Multi-level memory for task oriented dialogs," *ArXiv*, vol. abs/1810.10647, 2018. 4, 21

[22] A. Madotto, C.-S. Wu, and P. Fung, "Mem2seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems," *ArXiv*, vol. abs/1804.08217, 2018. 4, 15

[23] Z. Zhang, M. Huang, Z. Zhao, F. Ji, H. Chen, and X. Zhu, "Memory-augmented dialogue management for task-oriented dialogue systems," *ACM Transactions on Information Systems (TOIS)*, vol. 37, pp. 1 – 30, 2018. 4, 7, 15, 18

[24] A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, "Self-rag: Learning to retrieve, generate, and critique through self-reflection," *ArXiv*, vol. abs/2310.11511, 2023. 4, 15

[25] H. Shi, Z. Xu, H. Wang, W. Qin, W. Wang, Y. Wang, and H. Wang, "Continual learning of large language models: A comprehensive survey," *ArXiv*, vol. abs/2404.16789, 2024. 5, 8

[26] M. Li, F. Song, Y. Bowen, H. Yu, Z. Li, F. Huang, and Y. Li, "Api-bank: A comprehensive benchmark for tool-augmented llms," in *Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 3102–3116. 5

[27] D. Zhang, Y. Yu, C. Li, J. Dong, D. Su, C. Chu, and D. Yu, "Mm-llms: Recent advances in multimodal large language models," in *Annual Meeting of the Association for Computational Linguistics*, 2024, pp. 12 401–12 430. 5, 20

[28] Y. Chen, Z. Wen, G. Fan, Z. Chen, W. Wu, D. Liu, Z. Li, B. Liu, and Y. Xiao, "Mapo: Boosting large language model performance with model-adaptive prompt optimization," in *Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 3279–3304. 5

[29] B. Peng, Z. Tian, S. Liu, M. Yang, and J. Jia, "Scalable language model with generalized continual learning," *ArXiv*, vol. abs/2404.07470, 2024. 5

[30] X. Liu, J. Wang, J. Sun, X. Yuan, G. Dong, P. Di, W. Wang, and D. Wang, "Prompting frameworks for large language models: A survey," *ArXiv*, vol. abs/2311.12785, 2023. 5

[31] L. Friedman, S. Ahuja, D. Allen, Z. Tan, H. Sidahmed, C. Long, J. Xie, G. Schubiner, A. Patel, H. Lara, B. Chu, Z. Chen, and M. Tiwari, "Leveraging large language models in conversational recommender systems," *ArXiv*, vol. abs/2305.07961, 2023. 5

[32] H. Joko, S. Chatterjee, A. Ramsay, A. D. Vries, J. Dalton, and F. Hasibi, "Doing personal laps: Llm-augmented dialogue construction for personalized multi-session conversational search," *ArXiv*, vol. abs/2405.03480, 2024. 5

[33] L. Sun, X. Chen, L. Chen, T. Dai, Z. Zhu, and K. Yu, "Meta-gui: Towards multi-modal conversational agents on mobile gui," in *Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 6699–6712. 6

[34] A. Madotto, Z. Lin, Y. Bang, and P. Fung, "The adapter-bot: All-in-one controllable conversational model," in *AAAI Conference on Artificial Intelligence*, 2020, pp. 16 081–16 083. 6

[35] H. Wen, Y. Liu, W. Che, L. Qin, and T. Liu, "Sequence-to-sequence learning for task-oriented dialogue with dialogue state representation," *ArXiv*, vol. abs/1806.04441, 2018. 6

[36] R. Xu, C. Tao, D. Jiang, X. Zhao, D. Zhao, and R. Yan, "Learning an effective context-response matching model with self-supervised tasks for retrieval-based dialogues," in *AAAI Conference on Artificial Intelligence*, 2020, pp. 14 158–14 166. 6

[37] A. Madotto and Z. Liu, "Language models as few-shot learner for task-oriented dialogue systems," *ArXiv*, vol. abs/2008.06239, 2020. 7

[38] P. Budzianowski and I. Vulic, "Hello, it's gpt-2 - how can i help you? towards the use of pretrained language models for task-oriented dialogue systems," *ArXiv*, vol. abs/1907.05774, 2019. 7, 11

[39] Z. Liu, M. Patwary, R. Prenger, S. Prabhumoye, W. Ping, M. Shoeybi, and B. Catanzaro, "Multi-stage prompting for knowledgeable dialogue generation," *ArXiv*, vol. abs/2203.08745, 2022. 7

[40] H. Zhong, Z. Dou, Y. Zhu, H. Qian, and J. rong Wen, "Less is more: Learning to refine dialogue history for personalized dialogue generation," *ArXiv*, vol. abs/2204.08128, 2022. 7

[41] C. Niu, X. Wang, X. Cheng, J. Song, and T. Zhang, "Enhancing dialogue state tracking models through llm-backed user-agents simulation," *ArXiv*, vol. abs/2405.13037, 2024. 7

[42] S. Zhang, L. Dong, X. Li, S. Zhang, X. Sun, S. Wang, J. Li, R. Hu, T. Zhang, F. Wu, and G. Wang, "Instruction tuning for large language models: A survey," *ArXiv*, vol. abs/2308.10792, 2023. 8

[43] R. Thoppilan, D. D. Freitas, J. Hall, N. M. Shazeer, A. Kulshreshtha, H.-T. Cheng, A. Jin, T. Bos, L. Baker, Y. Du, Y. Li, H. Lee, H. Zheng, A. Ghafouri, M. Menegali, Y. Huang, M. Krikun, D. Lepikhin, J. Qin, D. Chen, Y. Xu, Z. Chen, A. Roberts, M. Bosma, Y. Zhou, C.-C. Chang, I. Krivokon, W. Rusch, M. Pickett, K. Meier-Hellstern, M. Morris, T. Doshi, R. D. Santos, T. Duke, J. Søraker, B. Zevenbergen, V. Prabhakaran, M. Díaz, B. Hutchinson, K. Olson, A. Molina, E. Hoffman-John, J. Lee, L. Aroyo, R. Rajakumar, A. Butryna, M. Lamm, V. Kuzmina, J. Fenton, A. Cohen, R. Bernstein, R. Kurzweil, B. Aguera-Arcas, C. Cui, M. Croak, E. Chi, and Q. Le, "Lamda: Language models for dialog applications," *ArXiv*, vol. abs/2201.08239, 2022. 8

[44] A. Kopf, Y. Kilcher, D. von Rutte, S. Anagnostidis, Z. R. Tam, K. Stevens, A. Barhoum, N. M. Duc, O. Stanley, R. Nagyfi, E. Shahul, S. Suri, D. Glushkov, A. Dantuluri, A. Maguire, C. Schuhmann, H. Nguyen, and A. Mattick, "Openassistant conversations - democratizing large language model alignment," *ArXiv*, vol. abs/2304.07327, 2023. 8, 20, 21

[45] S. Gao, Y. Zhang, Z. Ou, and Z. Yu, "Paraphrase augmented task-oriented dialog generation," *ArXiv*, vol. abs/2004.07462, 2020. 9

[46] J. Ni, T. Young, V. Pandelea, F. Xue, V. Adiga, and E. Cambria, "Recent advances in deep learning based dialogue systems: a systematic survey," *Artificial Intelligence Review*, vol. 56, pp. 3055–3155, 2021. 9, 12, 17, 20

[47] X. Li, Y.-N. V. Chen, L. Li, J. Gao, and A. Celikyilmaz, "End-to-end task-completion neural dialogue systems," *ArXiv*, vol. abs/1703.01008, 2017. 10, 21

[48] H. Chen, X. Liu, D. Yin, and J. Tang, "A survey on dialogue systems: Recent advances and new frontiers," *ArXiv*, vol. abs/1711.01731, 2017. 10, 13, 16, 20

[49] L. Xu, J. Zhang, B. Li, J. Wang, M. Cai, W. X. Zhao, and J.-R. Wen, "Prompting large language models for recommender systems: A comprehensive framework and empirical analysis," *ArXiv*, vol. abs/2401.04997, 2024. 10

[50] H. Wang, L. Wang, Y. Du, L. Chen, J. Zhou, Y. Wang, and K.-F. Wong, "A survey of the evolution of language model-based dialogue systems," *ArXiv*, vol. abs/2311.16789, 2023. 10, 14

[51] Y. Shao, S. Gouws, D. Britz, A. Goldie, B. Strope, and R. Kurzweil, "Generating high-quality and informative conversation responses with sequence-to-sequence models," in *Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 2210–2219. 10

[52] A. Sanders, T. Strzalkowski, M. Si, A. L. S. Chang, D. Dey, J. Braasch, D. W. R. P. Institute, Troy, Ny, Usa, and I. Research, "Towards a progression-aware autonomous dialogue agent," *ArXiv*, vol. abs/2205.03692, 2022. 10

[53] I. Serban, A. Sordoni, R. Lowe, L. Charlin, J. Pineau, A. C. Courville, and Y. Bengio, "A hierarchical latent variable encoder-decoder model for generating dialogues," *ArXiv*, vol. abs/1605.06069, 2016. 10, 18

[54] T. Wolf, V. Sanh, J. Chaumond, and C. Delangue, "Transfertransfo: A transfer learning approach for neural network based conversational agents," *ArXiv*, vol. abs/1901.08149, 2019. 11, 18

[55] M. Song, M. Zheng, and X. Luo, "Counting-stars: A simple, efficient, and reasonable strategy for evaluating long-context large language models," *ArXiv*, vol. abs/2403.11802, 2024. 11, 18

[56] Z. Lin, A. Madotto, G. I. Winata, P. Xu, F. Jiang, Y. Hu, C. Shi, and P. Fung, "Bitod: A bilingual multi-domain dataset for task-oriented dialogue modeling," *ArXiv*, vol. abs/2106.02787, 2021. 11

[57] Y.-T. Lin and Y.-N. V. Chen, "Llm-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models," *ArXiv*, vol. abs/2305.13711, 2023. 11

[58] W.-C. Kwan, X. Zeng, Y. Jiang, Y. Wang, L. Li, L. Shang, X. Jiang, Q. Liu, and K.-F. Wong, "Mt-eval: A multi-turn capabilities evaluation benchmark for large language models," *ArXiv*, vol. abs/2401.16745, 2024. 12, 19

[59] V. M. Andreas, G. I. Winata, and A. Purwarianti, "A comparative study on language models for task-oriented dialogue systems," *2021 8th International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*, pp. 1–5, 2021. 12

[60] M. Wang, I. Shafran, H. Soltau, W. Han, Y. Cao, D. Yu, and L. E. Shafey, "Retrieval augmented end-to-end spoken dialog models," *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12 056–12 060, 2024. 12

[61] Y. Dai, H. Yu, Y. Jiang, C. Tang, Y. Li, and J. Sun, "A survey on dialog management: Recent advances and challenges," *ArXiv*, vol. abs/2005.02233, 2020. 12, 16, 20

[62] B. Liu, G. Tür, D. Z. Hakkani-Tür, P. Shah, and L. Heck, "Dialogue learning with human teaching and feedback in end-to-end trainable task-oriented dialogue systems," in *North American Chapter of the Association for Computational Linguistics*, 2018, pp. 2060–2069. 12, 16, 21

[63] Z. Zeng, J. Yu, T. Gao, Y. Meng, T. Goyal, and D. Chen, "Evaluating large language models at evaluating instruction following," *ArXiv*, vol. abs/2310.07641, 2023. 12

[64] B. Wang, J. Liu, J. Karimnazarov, and N. Thompson, "Task supportive and personalized human-large language model interaction: A user study," *Proceedings of the 2024 Conference on Human Information Interaction and Retrieval*, 2024. 13

[65] X. Feng, Z. Chen, Y. Qin, Y. Lin, X. Chen, Z. Liu, and J.-R. Wen, "Large language model-based human-agent collaboration for complex task solving," *ArXiv*, vol. abs/2402.12914, 2024. 13

[66] I. Gur, D. Z. Hakkani-Tür, G. Tür, and P. Shah, "User modeling for task oriented dialogues," *2018 IEEE Spoken Language Technology Workshop (SLT)*, pp. 900–906, 2018. 13, 17, 21

[67] P. Budzianowski, T.-H. Wen, B.-H. Tseng, I. Casanueva, S. Ultes, O. Ramadan, and M. Gasic, "Multiwoz - a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling," in *Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 5016–5026. 13, 17

[68] S. Ouyang, S. Wang, Y. Liu, M. Zhong, Y. Jiao, D. Iter, R. Pryzant, C. Zhu, H. Ji, and J. Han, "The shifted and the overlooked: A task-oriented investigation of user-gpt interactions," in *Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 2375–2393. 13

[69] Z. Zheng, Y. Wang, Y. Huang, S. Song, B. Tang, F. Xiong, and Z. Li, "Attention heads of large language models: A survey," *ArXiv*, vol. abs/2409.03752, 2024. 14

[70] Z. Yu, Z. Wang, Y. Fu, H. Shi, K. Shaikh, and Y. Lin, "Unveiling and harnessing hidden attention sinks: Enhancing large language models without training through attention calibration," *ArXiv*, vol. abs/2406.15765, 2024. 14

[71] G. Lee, V. Hartmann, J. Park, D. Papailiopoulos, and K. Lee, "Prompted llms as chatbot modules for long open-domain conversation," in *Annual Meeting of the Association for Computational Linguistics*, 2023, pp. 4536–4554. 14

[72] B. Gao, Z. He, P. Sharma, Q. Kang, D. Jevdjic, J. Deng, X. Yang, Z. Yu, and P. Zuo, "Cost-efficient large language model serving for multi-turn conversations with cachedattention," in *USENIX Annual Technical Conference*, 2024, pp. 111–126. 14

[73] J. Yang, H. Jin, R. Tang, X. Han, Q. Feng, H. Jiang, B. Yin, and X. Hu, "Harnessing the power of llms in practice: A survey on chatgpt and beyond," *ACM Transactions on Knowledge Discovery from Data*, vol. 18, pp. 1 – 32, 2023. 14

[74] Z. Lin, A. Madotto, G. I. Winata, and P. Fung, "Mintl: Minimalist transfer learning for task-oriented dialogue systems," *ArXiv*, vol. abs/2009.12005, 2020. 14

[75] F. Wan, X. Huang, D. Cai, X. Quan, W. Bi, and S. Shi, “Knowledge fusion of large language models,” *ArXiv*, vol. abs/2401.10491, 2024. 14

[76] Z. Zhang, Y. Gu, X. Han, S. Chen, C. Xiao, Z. Sun, Y. Yao, F. Qi, J. Guan, P. Ke, Y. Cai, G. Zeng, Z. Tan, Z. Liu, M. Huang, W. Han, Y. Liu, X. Zhu, and M. Sun, “Cpm-2: Large-scale cost-effective pre-trained language models,” *AI Open*, vol. 2, pp. 216–224, 2021. 15

[77] R. Zhang, Y. Li, Y. Ma, M. Zhou, and L. Zou, “Llmaaa: Making large language models as active annotators,” *ArXiv*, vol. abs/2310.19596, 2023. 15

[78] F. Mi, M. Huang, J. Zhang, and B. Faltings, “Meta-learning for low-resource natural language generation in task-oriented dialogue systems,” in *International Joint Conference on Artificial Intelligence*, 2019, pp. 3151–3157. 16

[79] B. Geng, F. Yuan, Q. Xu, Y. Shen, R. Xu, and M. Yang, “Continual learning for task-oriented dialogue system with iterative network pruning, expanding and masking,” in *Annual Meeting of the Association for Computational Linguistics*, 2021, pp. 517–523. 16

[80] S. Santhanam and S. Shaikh, “A survey of natural language generation techniques with a focus on dialogue systems - past, present and future directions,” *ArXiv*, vol. abs/1906.00500, 2019. 16

[81] S. Sharma, L. E. Asri, H. Schulz, and J. Zumer, “Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation,” *ArXiv*, vol. abs/1706.09799, 2017. 17

[82] C. Xing, W. Wu, Y. Wu, J. Liu, Y. Huang, M. Zhou, and W.-Y. Ma, “Topic aware neural response generation,” in *AAAI Conference on Artificial Intelligence*, 2016, pp. 3351–3357. 18

[83] T. Whang, D. Lee, D. Oh, C. Lee, K. Han, D. Lee, and S. Lee, “Do response selection models really know what’s next? utterance manipulation strategies for multi-turn response selection,” in *AAAI Conference on Artificial Intelligence*, 2020, pp. 14 041–14 049. 18

[84] X. Zhao, W. Wu, C. Xu, C. Tao, D. Zhao, and R. Yan, “Knowledge-grounded dialogue generation with pre-trained language models,” in *Conference on Empirical Methods in Natural Language Processing*, 2020, pp. 3377–3390. 18

[85] S. K. Freire, C. Wang, and E. Niforatos, “Conversational assistants in knowledge-intensive contexts: An evaluation of llm- versus intent-based systems,” 2024. 19

[86] Y. Deng, W. Lei, H. Wang, and T. seng Chua, “Prompting and evaluating large language models for proactive dialogues: Clarification, target-guided, and non-collaboration,” in *Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 10 602–10 621. 19

[87] J. Liu, J. Jin, Z. Wang, J. Cheng, Z. Dou, and J. rong Wen, “Reta-llm: A retrieval-augmented large language model toolkit,” *ArXiv*, vol. abs/2306.05212, 2023. 19

[88] M. Wang, L. Chen, C. Fu, S. Liao, X. Zhang, B. Wu, H. Yu, N. Xu, L. Zhang, R. Luo, Y. Li, M. Yang, F. Huang, and Y. Li, “Leave no document behind: Benchmarking long-context llms with extended multi-doc qa,” *ArXiv*, vol. abs/2406.17419, 2024. 19

[89] C.-S. Wu and C. Xiong, “Probing task-oriented dialogue representation from language models,” in *Conference on Empirical Methods in Natural Language Processing*, 2020, pp. 5036–5051. 20

[90] T. S. Wu, M. Terry, and C. J. Cai, “Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts,” *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2021. 20

[91] S. Maity, A. Deroy, and S. Sarkar, “Exploring the capabilities of prompted large language models in educational and assessment applications,” *ArXiv*, vol. abs/2405.11579, 2024. 20

[92] H. Shi, Z. Xu, H. Wang, W. Qin, W. Wang, Y. Wang, and H. Wang, “Continual learning of large language models: A comprehensive survey,” *ArXiv*, vol. abs/2404.16789, 2024. 20

[93] Z. Wang, S. X. Teo, J. Ouyang, Y. Xu, and W. Shi, “M-rag: Reinforcing large language model performance through retrieval-augmented generation with multiple partitions,” *ArXiv*, vol. abs/2405.16420, 2024. 20

[94] S. Kim, M. Eric, K. Gopalakrishnan, B. Hedayatnia, Y. Liu, and D. Z. Hakkani-Tür, “Beyond domain apis: Task-oriented conversational modeling with unstructured knowledge access,” *ArXiv*, vol. abs/2006.03533, 2020. 21

[95] Y. Wang, M. Wang, M. A. Manzoor, F. Liu, G. Georgiev, R. J. Das, and P. Nakov, “Factuality of large language models in the year 2024,” *ArXiv*, vol. abs/2402.02420, 2024. 21

[96] M. Gao, X. Hu, J. Ruan, X. Pu, and X. Wan, “Llm-based nlg evaluation: Current status and challenges,” *ArXiv*, vol. abs/2402.01383, 2024. 21