Two Tales of Persona in LLMs: A Survey of Role-Playing and Personalization

Anonymous ACL submission

Abstract

The concept of *persona*, originally adopted in dialogue literature, has re-surged as a promising framework for tailoring large language models (LLMs) to specific context (*e.g.*, personalized search, LLM-as-a-judge). However, the growing research on leveraging persona in LLMs is relatively disorganized and lacks a systematic taxonomy. To close the gap, we present a comprehensive survey to categorize the current state of the field. We identify two lines of research, namely (1) *LLM Role-Playing*, where personas are assigned to LLMs, and (2) *LLM Personalization*, where LLMs take care of user personas. Additionally, we introduce existing methods for LLM personality evaluation. To the best of our knowledge, we present the first survey for role-playing and personalization in LLMs under the unified view of persona. We continuously maintain a paper collection to foster future endeavors.

1 Introduction

The striking capabilities of large language models (LLMs), exemplified by ChatGPT (OpenAI, 2022), have significantly advanced the field of natural language processing (NLP; Wei et al., 2023; Madaan et al., 2024; Shinn et al., 2024). Recently, in addition to using LLMs as NLP task solvers or general-purpose chatbots, the question of *how to adapt LLMs for specific context* has received great attention. To this end, leveraging *personas* has resurfaced as an ideal lens for adapting LLMs in target scenarios (Chen et al., 2023a, 2024). By incorporating personas, LLMs can generate more contextually appropriate responses, maximizing their utility and effectiveness for specific applications. However, the growing literature on persona in the LLM era is relatively disorganized, lacking a unifying overview.

In this paper, we aim to close the gap by offering a comprehensive survey and a systematic categorization of existing studies. Specifically, we divide current research into two main streams, namely *LLM Role-Playing* and *LLM Personalization*, as illustrated in Figure 1. The primary distinction is that in role-playing, the persona belongs to the LLM, while in personalization, the persona belongs to the user. The definitions are detailed below.

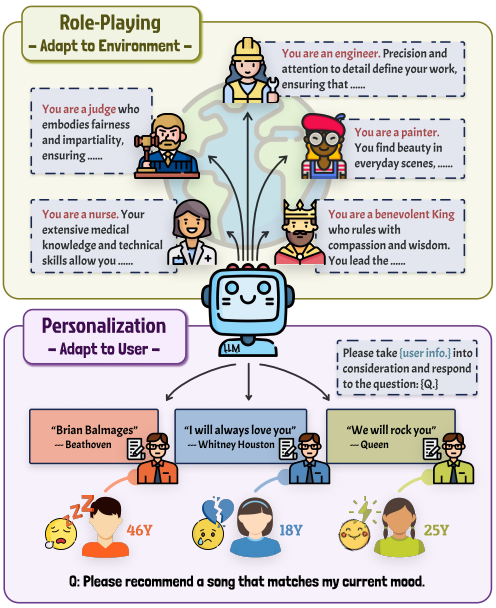

Figure 1: In *Role-Playing*, LLMs act according to assigned personas (*i.e.*, roles) under a defined environment. For example, given *role names* with *descriptions*, LLMs role-play in a social simulation game. For *Personalization*, LLMs consider *user personas* to generate tailored responses to the same question. Dashed rectangles are prompts and solid rectangles are LLMs' responses.

- ***LLM Role-Playing*:** LLMs are tasked to play the assigned personas (*i.e.,* roles) and act based on environmental feedback, adapting to the environment.

- ***LLM Personalization*:** LLMs are tasked to take care of user personas (*e.g.*, background information or historical behaviors) to meet individualized needs, adapting to distinct users.

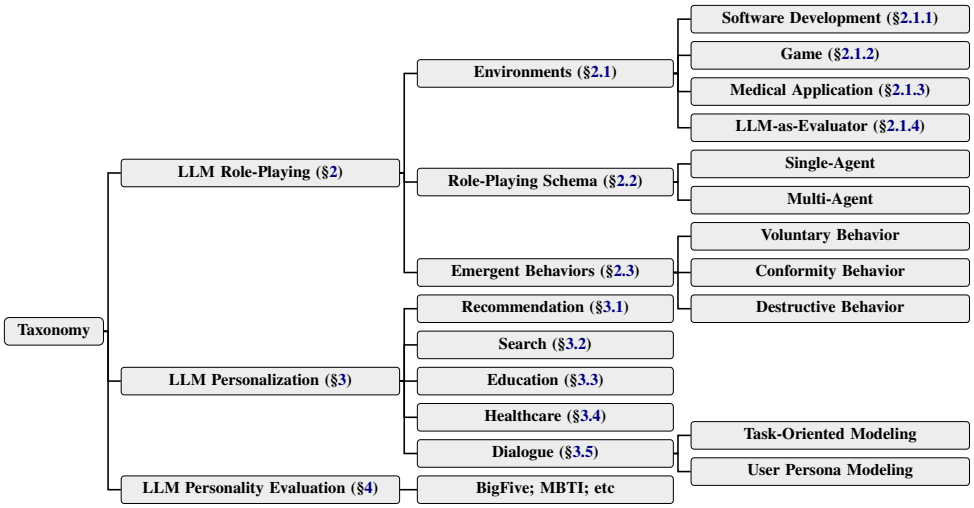

Figure 2: The taxonomy of LLM role-playing and LLM personalization.

To the best of our knowledge, we present the first survey for LLM role-playing and LLM personalization under the unified view of persona. To foster future endeavors, we actively maintain a paper collection available to the research community. We aim for this work to serve as both a valuable introduction for newcomers and a comprehensive resource for current researchers in the field.

Our taxonomy is illustrated in Figure 2. We first introduce LLM role-playing (§2), followed by LLM personalization (§3). Next, we provide an overview of evaluation methods (§4) assessing whether the personality of LLMs (*e.g.*, personality traits or psychological behaviors) accurately aligns with expected personas after the adaptation (*i.e.*, for role-playing LLMs that act according to assigned personas and personalized LLMs that fit user personas). Lastly, we highlight current challenges and future directions (§5). A comprehensive list of benchmarks and datasets is provided in the Appendix.

2 LLM Role-Playing

LLM-based language agents have demonstrated impressive abilities, such as planning, reflection, and tool-use recently (Yao et al., 2022b; Shinn et al., 2024; Yao et al., 2024). The predominant approach of LLM role-playing is by coupling personas with language agents, specifically, by incorporating personas directly inside the prompt of language agents. Such a training-free paradigm is particularly desirable due to its simplicity and effectiveness.

Language agents with role-playing elicit the corresponding parametric knowledge in LLMs to generate responses aligned with assigned personas (*i.e.*, role), enabling them to adapt to various interactive environments. LLM role-playing also extends to *multi-agent* settings, where multiple language agents are equipped with diverse personas, cooperating and communicating with each other to solve complex tasks (Guo et al., 2024). For instance, in one of the first works of role-played LLMs, Park et al. (2023) propose *generative agents*, which engage in a social simulation environment by mimicking human behaviors according to names, ages, and personality traits specified in the prompts.

Following we introduce different environments and associated roles in which LLMs adapt to (§2.1), interactions between LLMs within the environment (§2.2), and emergent behaviors stemming from their interactions (§2.3). Figure 3 provides an illustrative overview.

2.1 Environments

2.1.1 Software Development

For software development, the goal typically involves designing programs or coding projects. For instance, *"Create a snake game."* or *"Create a Python program to develop an interactive weather dashboard."* (Hong et al., 2023a). Due to the complexity of these tasks, often too intricate to be completed correctly on the first attempt, existing research leverages approaches like the Waterfall model (Petersen et al., 2009; Bassil, 2012) or Standardized Operating Procedures (SOPs) (Belbin and Brown, 2022; DeMarco and Lister, 2013) to break down the tasks into manageable sub-tasks.

Similar to real-world settings, LLMs role-play to operate as a company in a collaborative, multi-agent software development environment (Qian

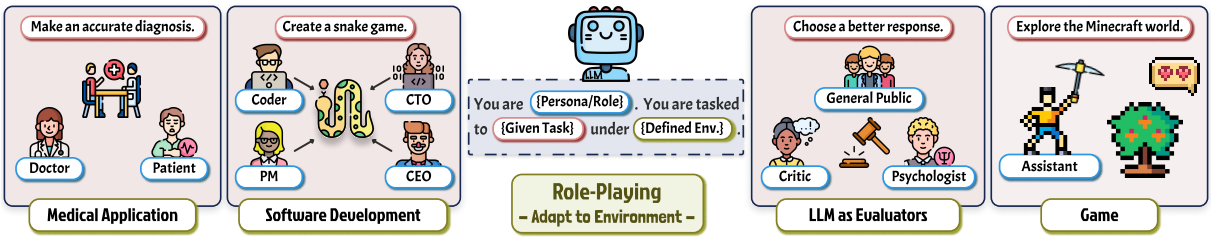

Figure 3: An illustration of five LLM role-playing environments: *Software Development* (§2.1.1), *Game* (§2.1.2), *Medical Application* (§2.1.3), and *LLM as Evaluators* (§2.1.4). For each environment, we provide a simple scenario with a task description (*red-bordered*) and relevant personas (*i.e.*, roles; *blue-bordered*). The dashed rectangle represents an example LLM role-playing prompt template. In addition to the above environments, past research also proposes general frameworks applicable to different environments (§5.1).

et al., 2023; Hong et al., 2023a; Dong et al., 2023). Different roles include Chief Technology Officer (CTO), Chief Product Officer (CPO), Chief Executive Officer (CEO), Product Managers, Engineers, Reviewers, and Testers. By assigning specific roles, LLMs are capable of carrying out tasks in a step-by-step and accurate manner.

Recent work (Dong et al., 2023) proposed one of the first self-collaboration frameworks that encompasses division of labor and collaboration among multiple LLM agents, each acting as a specialized *"experts"* to address complex code generation tasks. Following the Waterfall model, ChatDev (Qian et al., 2023) divides the development process into a four-phase pipeline: designing, coding, testing, and documenting and proposes *Chat Chain* to decompose each phase into a sequence of atomic sub-tasks. Differing from the above work, MetaGPT (Hong et al., 2023a) require LLM agents to generate structured outputs instead of free-text, demonstrating a significant increase in the success rate of target code generation.

2.1.2 Game

LLMs have been an effective backbone for agents in a variety of game environments, including Minecraft (Wang et al., 2023a), social simulation (Park et al., 2023; Wang et al., 2023d), and bargaining game (Fu et al., 2023). In these environments, LLMs are tasked to role-play as a general assistant (Wang et al., 2023a), or characters related to the environment, such as buyers and sellers (Fu et al., 2023). Gaming environments usually contain a wide range of information, including settings, utilizable tools, and nearby situations, which presents challenges for LLMs to memorize and respond. Thus, retrieval-based memory stream approaches are a crucial component for the effectiveness of language agents role-playing in the game environments (Park et al., 2023; Wang et al., 2023a).

2.1.3 Medical Application

In medical domain environments, Wu et al. (2023a) propose DR-CoT prompting, the first approach to leverage LLM role-playing for diagnostic reasoning. By mimicking doctors underlying thought processes, DR-CoT exhibits a striking improvement from standard prompting. Then, Kwon et al. (2024) extend such success to image-based diagnosis via knowledge distillation, addressing the application in real-world clinical settings. Another work, MedAgent (Tang et al., 2023a), introduces a multi-agent collaboration framework into medical reasoning through five processes: expert gathering, analysis proposition, report summarization, collaborative consultation, and decision making, to mimic actual medical scenarios.

These studies assign medically relevant personas to LLMs, ranging from general roles like doctor and patient to specific ones such as neurology and psychiatry experts. Their research demonstrates LLMs inherently possess medical knowledge (Liévin et al., 2024), enabling performance enhancement via LLM role-playing successfully.

2.1.4 LLM-as-Evaluator

The concept of adopting strong LLMs as evaluators has become a de facto framework for evaluating LM alignment. It is shown that LLMs are capable of assessing human-like values in model responses, and judgments made by LLMs could reflect a higher correlation with human ground-truth than traditional metrics (Chiang and Lee, 2023; Wang et al., 2023b; Lin and Chen, 2023).

Aiming for a greater similarity with human evaluation, roles in LLM-as-evaluator environments span a broad spectrum, representing various perspectives of human beings in society, such as the general public, the critic, and the news author. In

LLM-as-a-judge (Zheng et al., 2023), LLMs role-play an impartial judge and consider factors such as helpfulness, relevance, accuracy, depth, and creativity. Wu et al. (2023b) propose DRPE to assess the quality of summarization by assigning LLMs statically objective roles and dynamically subjective roles based on task settings. Another work, ChatEval (Chan et al., 2023), further adds discussion rounds within roles to improve the evaluation process, simulating a judge group in reality.

2.2 Role-Playing Schema

We categorize two schemas in LLM role-playing environments: *single-agent* and *multi-agent*.

Single-Agent We define the single-agent schema as follows: One agent is able to achieve its goal independently without assistance from others, though multiple agents may coexist in the same environment.

Single-agent schema is most common in game environments, where LLMs attend more to environmental information and feedback rather than collaboration. For example, Voyager (Wang et al., 2023a) agents, playing general assistant roles, are tasked to continuously explore the defined environment, acquire diverse skills, and make novel discoveries in Minecraft. Despite the presence of multiple Voyager agents in Minecraft, each agent is capable of exploring the gaming world on its own.

Multi-Agent We define the multi-agent schema as follows: Supports (*e.g.*, collaborate and communicate) from other agents are necessary for one agent to achieve its goal.

Software development and medical applications are the primary environments for multi-agent schema. Similar to real world, interaction within environments is crucial. Representative works like AgentVerse (Chen et al., 2023c) and ChatDev (Qian et al., 2023) both propose multi-agent frameworks that exchange information and cooperate to accomplish their tasks efficiently. Further, we identify two collaboration paradigms in the multi-agent schema (Xi et al., 2023; Guo et al., 2024): *Cooperative* and *Adversarial*. The cooperative paradigm facilitates information sharing among agents, for example, several works use message pools to store each agent's current state and ongoing tasks (Hong et al., 2023a; Tang et al., 2023a; Chen et al., 2023c). For the adversarial paradigm, including debate, competition, and criticism, enhances the decision-making process and seeks more advantages by adopting opposing perspectives (Chan et al., 2023; Fu et al., 2023).

2.3 Emergent Behaviors in Role-Playing

Under the multi-agent schema, different behaviors reflecting phenomena in human society (*e.g.*, conformity and consensus reaching) emerge through LLM collaboration. We introduce three collaborative behaviors following Chen et al. (2023c).

Voluntary Behavior Voluntary behaviors usually occur in the cooperative collaboration paradigm, where agents proactively assist their peers or inquire if there is anything they can help with to accomplish team goals. In addition, they may contribute resources to others, such as unallocated time and possessed materials. Through voluntary behaviors, LLMs enhance team efficiency and demonstrate cohesion and commitment within defined environments (Chen et al., 2023c; Hong et al., 2023a).

Conformity Behavior Conformity behaviors occur in situations where an agent deviates from the team goal. After receiving criticism and suggestions from others, the deviating agent then refines and adjusts its behavior or decisions to better cooperate with the team. Through conformity behaviors, LLMs align with the mutual goal and pursue improved accuracy and completeness (Tang et al., 2023a; Fu et al., 2023).

Destructive Behavior Occasionally, LLMs undertake various actions that lead to undesired and detrimental outcomes. For instance, it may exhibit a *Bad Mind* that seeks to control the world (Li et al., 2024a). Additionally, LLMs might display toxicity or reveal deep-seated stereotypical biases when equipping personas (Deshpande et al., 2023; Gupta et al., 2023). Such destructive behaviors raise safety and bias concerns of role-playing.

3 LLM Personalization

Prominent approaches for aligning LLMs to user intents typically leverage reinforcement learning from human feedback (RLHF), a process that infuses collective consciousness and biases into the model. To enhance individual experience and preference, personalized LLMs consider user personas (*e.g.,* individual information, historical behaviors) and cater to customized needs (Chen et al., 2023e; Deshpande et al., 2024). Following we introduce various personalized tasks with associated methods

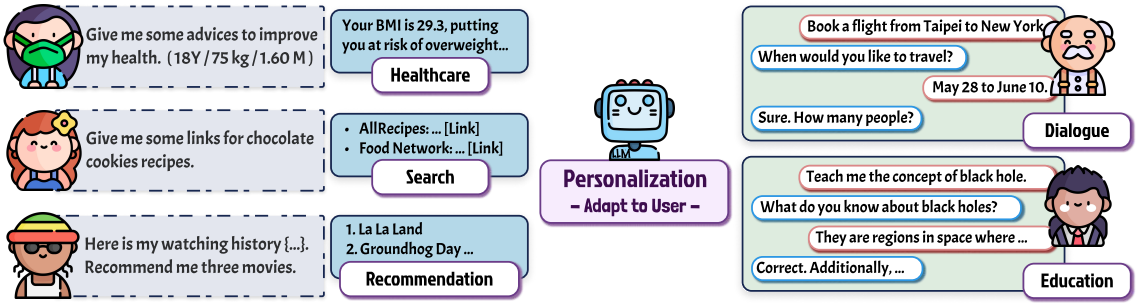

Figure 4: An illustration of five types of personalized LLMs: *Recommendation* (§3.1), *Search* (§3.2), *Education* (§3.3), *Healthcare* (§3.4), and *Dialogue* (§3.5). On the left side, dashed rectangles are prompts, and solid rectangles are the responses of LLMs. On the right side, we depict multi-turn interactions between LLMs and users.

for achieving personalization. Figure 4 presents an illustrative overview of personalization tasks.

3.1 Personalized Recommendation

Recommendation systems aim to recommend items (*e.g.*, books or movies) to users that match their preferences. We compare existing research in Table 3 and compile relevant datasets in Table 4.

Existing studies explore various prompting methods for using LLMs in recommendation systems. Li et al. (2023a) develop a method for efficient incorporation of users' personal information. Li et al. (2023b) combine aspect extraction with aspect-based recommendations via LLMs prompt tuning. Chen et al. (2022) generate personalized chit-chat to enhance recommendation. Focusing on the framework design, Yang et al. (2023b) present a novel LLM fine-tuning recommendation system. Chu et al. (2023) merge different recommendation systems to address the challenge of effectively integrating the commonsense and reasoning abilities of LLMs into recommendation systems. Hu et al. (2024) propose a sequential recommendation framework to preserve fine-grained item textual information.

A lot of works have focused on the zero-shot setting, leveraging the powerful out-of-the-box capabilities of LLMs. Wang and Lim (2023) adopt a three-step prompting pipeline to achieve better zero-shot next-item recommendation. Hou et al. (2024) propose a zero-shot sequential recommendation system via in-context learning. Zhang et al. (2023) enhance user-friendliness by allowing users to freely interact with the system and receive more precise recommendations through natural language instructions. For generalizability, Wang et al. (2024d) highlight that current recommendation systems mostly focus on specific tasks and lack the ability to generalize to new tasks. They propose an LLM-powered agent for general recommendation purposes. Although LLM-based personalized search systems present a more convenient and simple solution for information search, ensuring the accountability and trustworthiness of the synthesized results still requires further development (Li et al., 2024b).

3.2 Personalized Search

Compared to traditional search systems that provide a list of hard-to-organize relevant results and are limited to simple queries, personalized search systems enable understanding of complex queries and past interactions to infer user preferences, synthesizing information from multiple sources and presenting it in a cohesive, natural language form.

Spatharioti et al. (2023) demonstrate that LLM-based search systems improve users' performance in certain situations. Ziems et al. (2023) suggest that LLMs act as built-in search engines given few-shot demonstrations. Specifically, LLMs can generate correct web URLs for corresponding documents. Building upon Zhou et al. (2021), Zhou et al. (2024) present a strategy to combine the cognitive memory mechanism with LLMs for personalized search, enabling LLMs to efficiently retrieve memory. Some works also leverage search engine results to enhance LLM personalization (Baek et al., 2024; Salemi and Zamani, 2024). Empirically, Sharma et al. (2024) conduct experiments to investigate how LLM-powered search systems could lead to opinion polarization.

3.3 Personalized Education

The capability of LLMs can be utilized in a variety of ways to facilitate personalized education. For example, LLMs can provide detailed, step-by-step explanations in the Socratic teaching style (Hao et al., 2024), answer questions on technical and

complicated subjects (Arefeen et al., 2023), and automatically summarize lectures to enhance learning experience (Gonzalez et al., 2023).

Personalized LLMs have the potential to create a more inclusive and equitable educational ecosystem, obviating the need for individuals to pay disproportionate fees. Recent works have illustrated various opportunities and visions for integrating LLMs into educational environments. These applications range from personalized learning and teaching assistance to homework assessment and feedback (Kasneci et al., 2023; Gan et al., 2023; Wang et al., 2024b; Jeon and Lee, 2023; Huber et al., 2024).

For example, EDUCHAT (Dan et al., 2023) pretrained models on an educational corpus to establish a foundational knowledge base, and subsequently fine-tune models on personalized tasks such as essay assessment, Socratic teaching, or emotional support. HUMSUM (Shehata et al., 2023) summarize personalized lecture transcripts from diverse scenarios, considering factors such as length, depth, tone, and complexity. This is followed by prompt tuning to modify the summary based on the personalization options given by users. Park et al. (2024) incorporate the student's affective state, cognitive state, and learning style into the prompt to create a personalized conversation-based tutoring system.

3.4 Personalized Healthcare

LLMs have exhibited expert-level capabilities in a range of general biomedical tasks, with the potential to integrate into people's everyday lives (Cohan et al., 2020; Milne-Ives et al., 2020; Singhal et al., 2023; Saab et al., 2024; Abbasian et al., 2024b).

Towards personalized healthcare assistant, Abbasian et al. (2024a) propose OPENCHA, an LLM agentic framework that integrates external data and personalized health data to address personalized medical problems. Following OPENCHA, Abbasian et al. (2024c) infuse domain-specific knowledge to effectively utilize health data, knowledge bases, and analytical tools for diabetes-related questions. MALP (Zhang et al., 2024a) combine parameter-efficient fine-tuning (PEFT) with a memory retrieval module to generate personalized medical responses. Other frameworks such as HEALTH-LLM (Jin et al., 2024b) combine LlamaIndex (Liu, 2022) to make diagnosis predictions, and is capable of generating personalized medical advice based on symptom descriptions provided by users. Moreover, LLMs also show great potential for psychotherapy (Stade et al., 2024; Chen et al., 2023b; Xu et al., 2024).

3.5 Personalized Dialogue Generation

Depending on the goals, dialogue generation tasks can be categorized into: (1) Task-oriented dialogue modeling (ToD modeling) and (2) User persona modeling. Following we discuss ToD modeling and User persona. We also organize various datasets for dialogue generation in Table 2.

ToD Modeling ToD modeling guides users in completing specific tasks, such as hotel bookings or restaurant reservations, through multiple interactive steps. See an example in Table 5.

Hudeček and Dusek (2023) leverage instruction-tuned LLMs and employ in-context learning for retrieval, and state tracking. Focusing on factuality, REFGPT (Yang et al., 2023a) generate truthful responses by augmenting the dialogue history with reliable sources and use prompts to guide LLM according to predefined dialogue settings. Li et al. (2024c); Hu et al. (2023) explore prompt augmentations; on the other hand, DSP (Li et al., 2024c) train a small policy model to generate hints and guide LLMs in completing tasks. A lot of works used LLMs to generate multi-turn dialogue as training datasets (Yang et al., 2023a; Huryn et al., 2022; Xu et al., 2023). Further, personalized dialogues have been applied in procedural content generation for customized dialogue generation in video games (Ashby et al., 2023).

User Persona Modeling User persona modeling detects the user persona based on dialogue history and generates customized responses tailored for each user. See an example in Appendix B.

COBERT (Zhong et al., 2020) proposed persona-based empathetic conversations using BERT with a two-hop co-attention mechanism (Lu et al., 2017) to refine embeddings and identify the most relevant response given the context and persona information. Song et al. (2020) utilized natural language inference (NLI) as an RL task with response persona as the reward to generate persona-consistent dialogue. Liu et al. (2020) proposed $\mathcal{P}^2$, a mutual persona perception model, and employ supervised training and self-play fine-tuning in the training process. Tang et al. (2023b) combined sparse persona descriptions, dense persona descriptions, and dialogue history to generate personalized responses.

4 LLM Personality Evaluation

In the previous sections, we summarize the current progress in LLM role-playing and LLM personalization. Equally important is the evaluation of whether the personality of LLMs accurately reflects the intended persona after the adaptation (*i.e.*, for role-playing LLMs that act based on designated personas and personalized LLMs tailored to individualized personas).

A line of works has carried out the evaluation leveraging human personality assessments, including BigFive (Jiang et al., 2023; Sorokovikova et al., 2024) and MBTI (Pan and Zeng, 2023; Song et al., 2024). For example, Sorokovikova et al. (2024); Jiang et al. (2024) quantitatively evaluate LLM personality based on the BigFive Personality Inventory (BFI) test and story writing test. In the BFI evaluation, LLMs often can reflect their intended persona accurately. Moreover, their personas often influence their linguistic style and personality consistency (Frisch and Giulianelli, 2024; Jiang et al., 2023). While most works focus solely on either semantic accuracy or personality consistency, Harrison et al. (2019) further explore controlling the two aspects simultaneously.

Jiang et al. (2024) introduce Machine Personality Inventory (MPI) for evaluating LLMs' personality traits. They use BigFive Personality Factors to evaluate each personality trait consisting of a series of descriptions and a set of options and statistically measure each trait. By comparing with human evaluation, they find that the internal consistency correlates with model capabilities. On the other hand, Pan and Zeng (2023) evaluate LLMs with the MBTI test to assess whether LLMs possess human-like personalities, and conclude that different LLMs have different MBTI types, which are often attributable to their training corpus. Moreover, they find that simply modifying the prompts is unlikely to change the MBTI type of LLMs.

Another work by Wang et al. (2024c) evaluate the personality fidelity of role-playing LLMs via personality test interviewing, and ask LLM to rate the score of each personality dimension according to the interview. Their results suggest that LLMs' demonstrated personalities align well with the assigned character personas. However, whether the aforementioned human psychometric tests are directly transferable to be applied to LLMs remains an open question (Dorner et al., 2023).

5 Limitations and Future Directions

5.1 Towards a General Framework

Despite the effectiveness of various role-playing frameworks, they are mostly task dependent and heavily rely on human-crafted personas. Both require prior knowledge and deep understanding of the tasks (Chen et al., 2023c). Consequently, enhancing the generalizability of the framework and employing automatic prompt engineering is a fruitful directions (Li et al., 2024a; Wang et al., 2023c).

To this end, Li et al. (2024a) propose a novel task-independent framework that allows agents to collaborate autonomously, but is limited to two roles and still requires human assigned personas. Subsequently, Wang et al. (2023c) introduce methods for LLMs to automatically identify personas based on given problems. Another work by Chen et al. (2023c) also enable LLMs to dynamically adjust the personas. However, they require prior knowledge of the intended tasks and pre-defined configuration (*e.g.*, the number of agents).

5.2 Long-Context Personas

Richardson et al. (2023) note that incorporating user history data into the prompt for personalizing LLMs could lead to input exceeding context length as well as increased inference costs. Leveraging retrieval-based methods may have the problem of potential information loss. Some works have proposed to summarize user profiles, design long-term memory mechanisms focusing on user portrait, prestoring user information, or ways to efficiently represent for retrieval augmentation (Richardson et al., 2023; Zhong et al., 2024; Zhang et al., 2024b; Sun et al., 2024). However, retrieval augmentation might be underperforming due to unrelated or noisy prompts (Tan et al., 2024). How to better store, encode, and integrate long-context personas in LLMs requires further investigation.

5.3 Lack of Datasets and Benchmarks

For LLM role-playing, several tasks lack suitable datasets with specific formats and environmental information (*e.g.*, game environments require information about configurations and tools). For personalized dialogue generation, user persona modeling lacks contradictory persona datasets that would more accurately represent real human behaviors (Kim et al., 2024b). Furthermore, LLM personalization faces a scarcity of high-quality personal data for model development due to privacy

concerns, hindering a thorough evaluation of different personalization methods. In addition, existing benchmarks for both LLM role-playing and personalization are relatively limited, lacking comprehensive evaluations across various dimensions (Chang et al., 2023). Therefore, expanding datasets and benchmarks for specialized environments and personal information under privacy protection is an important future direction.

5.4 Bias

While a large number of studies focus on enhancing end-task performance, fewer works explore the biases induced by role-playing and personlization in LLMs. In this context, Gupta et al. (2023), as one of the first studies, highlights the deep-seated stereotypical biases found in LLMs assigned with socio-demographic personas. For personalized LLM recommendation systems, biases can be observed due to item popularity or item positions in the prompts (Hou et al., 2024). Empirically, Dorner et al. (2023) also reveal the presence of *agree bias* in LLMs – a tendency to agree with both true and false content, regardless of the actual facts. In sum, there exists ample room for investigating and mitigating different classes of biases in the context of LLM role-playing and personalization.

5.5 Safety and Privacy

Past research has shown safety issues in LLM role-playing and personalization. Jin et al. (2024a) and Shah et al. (2023) successfully manipulate LLMs to perform jailbreak collaboratively. Deshpande et al. (2023) also show that assigning personas to LLMs aid in jailbreaking. Negative behaviors in LLM role-playing are also demonstrated by Chen et al. (2023c) and Li et al. (2024a). Further, Deshpande et al. (2023) find that LLMs consistently exhibit toxicity in a range of topics when assigned personas. These works demonstrate the discovery of unsafe problems, indicating an urgent need and more efforts to prevent potential exploits.

Since LLM personalization heavily relies on user personas, including personal information and historical behaviors, ensuring privacy is especially crucial. Recently, Wang et al. (2024a) discover that using the membership inference attack can leak personal information, raising concerns about encoding personal data into models. Although existing research provides methods to address this personal information leakage (Lukas et al., 2023; Gambarelli et al., 2023; Huang et al., 2022; Chen et al., 2023d), the risks remain in need of more effort and attention from the research community.

5.6 Broader Influences

As LLM personalization continues to advance in education domains, individuals could easily access personalized educational contents, lecture materials, and receive affordable tutoring, largely benefiting minority groups with limited resources. However, the concern of polarizing trends might arise, where the privileged group enjoys private tutors and underrepresented individuals only have access to LLM-powered supports (Li et al., 2023c). Also, personalized LLMs for healthcare could potentially be widely integrated into clinical scenarios, mental health assessments, or prescribed therapeutic treatments in the near future, where critical questions such as legal considerations of the liability associated with these personalized systems needs careful considerations (Swift and Allen, 2010).

As discussed in (§4), though methods for LLM personality evaluation have been proposed, there still lacks a unifying understanding of how to quantify personality in LLMs (Fang et al., 2023). Song et al. (2024); Jiang et al. (2024) also show that LLMs sometimes do not hold consistent personalities. It is crucial to continuously explore new measurements for reliable assessment of personality and psychological traits in LLMs, considering that in the future they might take on more advanced roles and capabilities in society.

6 Conclusion

Leveraging personas, LLMs can generate tailored responses and effectively adapt to a wide range of scenarios. In this survey paper, we summarize two lines of work – role-playing and personalization – for research of personas in the era of LLMs. We also present various evaluation methods for LLM personality. Lastly, we highlight current challenges and promising future directions. We hope our extensive survey and resources serve as an introductory guide for beginners to the field and a practical roadmap to foster future endeavors.

References

Mahyar Abbasian, Iman Azimi, Amir M. Rahmani, and Ramesh Jain. 2024a. Conversational health agents: A personalized llm-powered agent framework. 6

Mahyar Abbasian, Elahe Khatibi, Iman Azimi, David Oniani, Zahra Shakeri Hossein Abad, Alexander

Thieme, Ram Sriram, Zhongqi Yang, Yanshan Wang, Bryant Lin, Olivier Gevaert, Li-Jia Li, Ramesh Jain, and Amir M. Rahmani. 2024b. Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative ai. 6

Mahyar Abbasian, Zhongqi Yang, Elahe Khatibi, Pengfei Zhang, Nitish Nagesh, Iman Azimi, Ramesh Jain, and Amir M. Rahmani. 2024c. Knowledge-infused llm-powered conversational health agent: A case study for diabetes patients. 6

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*. 17

Md Adnan Arefeen, Biplob Debnath, and Srimat Chakradhar. 2023. Leancontext: Cost-efficient domain-specific question answering using llms. 6

Trevor Ashby, Braden K Webb, Gregory Knapp, Jackson Searle, and Nancy Fulda. 2023. Personalized quest and dialogue generation in role-playing games: A knowledge graph-and language model-based approach. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–20. 6

Jinheon Baek, Nirupama Chandrasekaran, Silviu Cucerzan, Allen herring, and Sujay Kumar Jauhar. 2024. Knowledge-augmented large language models for personalized contextual query suggestion. 5

Youssef Bassil. 2012. A simulation model for the waterfall software development life cycle. *arXiv preprint arXiv:1205.6904*. 2

R Meredith Belbin and Victoria Brown. 2022. *Team roles at work*. Routledge. 2

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics. 18

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*. 4

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2023. A survey on evaluation of large language models. 8

Changyu Chen, Xiting Wang, Xiaoyuan Yi, Fangzhao Wu, Xing Xie, and Rui Yan. 2022. Personalized chit-chat generation for recommendation using external chat corpora. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 2721–2731. 5, 19

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, et al. 2024. From persona to personalization: A survey on role-playing language agents. *arXiv preprint arXiv:2404.18231*. 1

Jin Chen, Zheng Liu, Xu Huang, Chenwang Wu, Qi Liu, Gangwei Jiang, Yuanhao Pu, Yuxuan Lei, Xiaolong Chen, Xingmei Wang, Defu Lian, and Enhong Chen. 2023a. When large language models meet personalization: Perspectives of challenges and opportunities. 1

Siyuan Chen, Mengyue Wu, Kenny Q Zhu, Kunyao Lan, Zhiling Zhang, and Lyuchun Cui. 2023b. Llm-empowered chatbots for psychiatrist and patient simulation: application and evaluation. *arXiv preprint arXiv:2305.13614*. 6

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chen Qian, Chi-Min Chan, Yujia Qin, Yaxi Lu, Ruobing Xie, et al. 2023c. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors in agents. *arXiv preprint arXiv:2308.10848*. 4, 7, 8

Yang Chen, Ethan Mendes, Sauvik Das, Wei Xu, and Alan Ritter. 2023d. Can language models be instructed to protect personal information? 8

Zheng Chen, Ziyan Jiang, Fan Yang, Zhankui He, Yupeng Hou, Eunah Cho, Julian McAuley, Aram Galstyan, Xiaohua Hu, and Jie Yang. 2023e. The first workshop on personalized generative ai@ cikm 2023: Personalization meets large language models. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 5267–5270. 4

Kanzhi Cheng, Qiushi Sun, Yougang Chu, Fangzhi Xu, Yantao Li, Jianbing Zhang, and Zhiyong Wu. 2024. Seeclick: Harnessing gui grounding for advanced visual gui agents. *arXiv preprint arXiv:2401.10935*. 17

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*. 3

Zhixuan Chu, Hongyan Hao, Xin Ouyang, Simeng Wang, Yan Wang, Yue Shen, Jinjie Gu, Qing Cui, Longfei Li, Siqiao Xue, James Y Zhang, and Sheng Li. 2023. Leveraging large language models for pre-trained recommender systems. 5, 19

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings*

of the 58th Annual Meeting of the Association for Computational Linguistics, pages 2270–2282, Online. Association for Computational Linguistics. 6

Yuhao Dan, Zhikai Lei, Yiyang Gu, Yong Li, Jianghao Yin, Jiaju Lin, Linhao Ye, Zhiyan Tie, Yougen Zhou, Yilei Wang, et al. 2023. Educhat: A large-scale language model-based chatbot system for intelligent education. *arXiv preprint arXiv:2308.02773*. 6

Tom DeMarco and Tim Lister. 2013. *Peopleware: productive projects and teams*. Addison-Wesley. 2

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Sam Stevens, Boshi Wang, Huan Sun, and Yu Su. 2024. Mind2web: Towards a generalist agent for the web. *Advances in Neural Information Processing Systems*, 36. 17, 18

Ameet Deshpande, EunJeong Hwang, Vishvak Murahari, Joon Sung Park, Diyi Yang, Ashish Sabharwal, Karthik Narasimhan, and Ashwin Kalyan, editors. 2024. *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*. Association for Computational Linguistics, St. Julians, Malta. 4

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. 4, 8

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, Shrimai Prabhumoye, Alan W Black, Alexander Rudnicky, Jason Williams, Joelle Pineau, Mikhail Burtsev, and Jason Weston. 2019. The second conversational intelligence challenge (convai2). 18

Yihong Dong, Xue Jiang, Zhi Jin, and Ge Li. 2023. Self-collaboration code generation via chatgpt. *arXiv preprint arXiv:2304.07590*. 3

Florian E Dorner, Tom Sühr, Samira Samadi, and Augustin Kelava. 2023. Do personality tests generalize to large language models? *arXiv preprint arXiv:2311.05297*. 7, 8

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur. 2020. MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 422–428, Marseille, France. European Language Resources Association. 18

Qixiang Fang, Anastasia Giachanou, Ayoub Bagheri, Laura Boeschoten, Erik-Jan van Kesteren, Mahdi Shafiee Kamalabad, and Daniel Oberski. 2023. On text-based personality computing: Challenges and future directions. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10861–10879, Toronto, Canada. Association for Computational Linguistics. 8

Ivar Frisch and Mario Giulianelli. 2024. LLM agents in interaction: Measuring personality consistency and linguistic alignment in interacting populations of large language models. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, pages 102–111, St. Julians, Malta. Association for Computational Linguistics. 7

Yao Fu, Hao Peng, Tushar Khot, and Mirella Lapata. 2023. Improving language model negotiation with self-play and in-context learning from ai feedback. *arXiv preprint arXiv:2305.10142*. 3, 4

Gaia Gambarelli, Aldo Gangemi, and Rocco Tripodi. 2023. Is your model sensitive? spedac: A new resource for the automatic classification of sensitive personal data. *IEEE Access*, 11:10864–10880. 8

Wensheng Gan, Zhenlian Qi, Jiayang Wu, and Jerry Chun-Wei Lin. 2023. Large language models in education: Vision and opportunities. In *2023 IEEE International Conference on Big Data (BigData)*, pages 4776–4785. IEEE. 6

Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2022. Recommendation as language processing (rlp): A unified pretrain, personalized prompt & predict paradigm (p5). In *Proceedings of the 16th ACM Conference on Recommender Systems*, pages 299–315. 19

Hannah Gonzalez, Jiening Li, Helen Jin, Jiaxuan Ren, Hongyu Zhang, Ayotomiwa Akinyele, Adrian Wang, Eleni Miltsakaki, Ryan Baker, and Chris Callison-Burch. 2023. Automatically generated summaries of video lectures may enhance students' learning experience. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 382–393, Toronto, Canada. Association for Computational Linguistics. 6

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*. 2, 4

Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. 2023. Bias runs deep: Implicit reasoning biases in persona-assigned llms. *arXiv preprint arXiv:2311.04892*. 4, 8

Izzeddin Gur, Hiroki Furuta, Austin Huang, Mustafa Safdari, Yutaka Matsuo, Douglas Eck, and Aleksandra Faust. 2023. A real-world webagent with planning, long context understanding, and program synthesis. *arXiv preprint arXiv:2307.12856*. 17

Izzeddin Gur, Ofir Nachum, Yingjie Miao, Mustafa Safdari, Austin Huang, Aakanksha Chowdhery, Sharan Narang, Noah Fiedel, and Aleksandra Faust. 2022. Understanding html with large language models. *arXiv preprint arXiv:2210.03945*. 17

Ji-Eun Han, Jun-Seok Koh, Hyeon-Tae Seo, Du-Seong Chang, and Kyung-Ah Sohn. 2024. Psydial: Personality-based synthetic dialogue generation using large language models. 18

Shibo Hao, Yi Gu, Haotian Luo, Tianyang Liu, Xiyan Shao, Xinyuan Wang, Shuhua Xie, Haodi Ma, Adithya Samavedhi, Qiyue Gao, Zhen Wang, and Zhiting Hu. 2024. Llm reasoners: New evaluation, library, and analysis of step-by-step reasoning with large language models. 5

F Maxwell Harper and Joseph A Konstan. 2015. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19. 19

Vrindavan Harrison, Lena Reed, Shereen Oraby, and Marilyn Walker. 2019. Maximizing stylistic control and semantic accuracy in NLG: Personality variation and discourse contrast. In *Proceedings of the 1st Workshop on Discourse Structure in Neural NLG*, pages 1–12, Tokyo, Japan. Association for Computational Linguistics. 7

Wanwei He, Yinpei Dai, Yinhe Zheng, Yuchuan Wu, Zheng Cao, Dermot Liu, Peng Jiang, Min Yang, Fei Huang, Luo Si, Jian Sun, and Yongbin Li. 2022. Galaxy: A generative pre-trained model for task-oriented dialog with semi-supervised learning and explicit policy injection. 18

Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, et al. 2023a. Metagpt: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*. 2, 3, 4

Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, et al. 2023b. Cogagent: A visual language model for gui agents. *arXiv preprint arXiv:2312.08914*. 17

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. *Advances in Neural Information Processing Systems*, 33:20179–20191. 17

Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. 2024. Large language models are zero-shot rankers for recommender systems. 5, 8, 19

Jun Hu, Wenwen Xia, Xiaolu Zhang, Chilin Fu, Weichang Wu, Zhaoxin Huan, Ang Li, Zuoli Tang, and Jun Zhou. 2024. Enhancing sequential recommendation via llm-based semantic embedding learning. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 103–111. 5

Zhiyuan Hu, Yue Feng, Yang Deng, Zekun Li, See-Kiong Ng, Anh Tuan Luu, and Bryan Hooi. 2023. Enhancing large language model induced task-oriented dialogue systems through look-forward motivated goals. 6

Jie Huang, Hanyin Shao, and Kevin Chen-Chuan Chang. 2022. Are large pre-trained language models leaking your personal information? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2038–2047, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. 8

Xinting Huang, Jianzhong Qi, Yu Sun, and Rui Zhang. 2020. Mala: Cross-domain dialogue generation with action learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7977–7984. 17

Stefan E Huber, Kristian Kiili, Steve Nebel, Richard M Ryan, Michael Sailer, and Manuel Ninaus. 2024. Leveraging the potential of large language models in education through playful and game-based learning. *Educational Psychology Review*, 36(1):1–20. 6

Vojtěch Hudeček and Ondrej Dusek. 2023. Are large language models all you need for task-oriented dialogue? In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 216–228, Prague, Czechia. Association for Computational Linguistics. 6

Daniil Huryn, William M. Hutsell, and Jinho D. Choi. 2022. Automatic generation of large-scale multi-turn dialogues from Reddit. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3360–3373, Gyeongju, Republic of Korea. International Committee on Computational Linguistics. 6

Léo Jacqmin, Lina M. Rojas-Barahona, and Benoit Favre. 2022. "do you follow me?": A survey of recent approaches in dialogue state tracking. 17

Jaeho Jeon and Seongyong Lee. 2023. Large language models in education: A focus on the complementary relationship between human teachers and chatgpt. *Education and Information Technologies*, 28(12):15873–15892. 6

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2024. Evaluating and inducing personality in pre-trained language models. *Advances in Neural Information Processing Systems*, 36. 7, 8

Hang Jiang, Xiajie Zhang, Xubo Cao, and Jad Kabbara. 2023. Personallm: Investigating the ability of large language models to express big five personality traits. *arXiv preprint arXiv:2305.02547*. 7

Haibo Jin, Ruoxi Chen, Andy Zhou, Jinyin Chen, Yang Zhang, and Haohan Wang. 2024a. Guard: Role-playing to generate natural-language jailbreakings to test guideline adherence of large language models. *arXiv preprint arXiv:2402.03299*. 8

Mingyu Jin, Qinkai Yu, Dong Shu, Chong Zhang, Lizhou Fan, Wenyue Hua, Suiyuan Zhu, Yanda Meng, Zhenting Wang, Mengnan Du, and Yongfeng Zhang. 2024b. Health-llm: Personalized retrieval-augmented disease prediction system. 6

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274. 6

Geunwoo Kim, Pierre Baldi, and Stephen McAleer. 2024a. Language models can solve computer tasks. *Advances in Neural Information Processing Systems*, 36. 17

Hana Kim, Kai Tzu-iunn Ong, Seoyeon Kim, Dongha Lee, and Jinyoung Yeo. 2024b. Commonsense-augmented memory construction and management in long-term conversations via context-aware persona refinement. *arXiv preprint arXiv:2401.14215*. 7

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*. 17

Jing Yu Koh, Robert Lo, Lawrence Jang, Vikram Duvvur, Ming Chong Lim, Po-Yu Huang, Graham Neubig, Shuyan Zhou, Ruslan Salakhutdinov, and Daniel Fried. 2024. Visualwebarena: Evaluating multimodal agents on realistic visual web tasks. *arXiv preprint arXiv:2401.13649*. 17, 18

Taeyoon Kwon, Kai Tzu-iunn Ong, Dongjin Kang, Seungjun Moon, Jeong Ryong Lee, Dosik Hwang, Beomseok Sohn, Yongsik Sim, Dongha Lee, and Jinyoung Yeo. 2024. Large language models are clinical reasoners: Reasoning-aware diagnosis framework with prompt-generated rationales. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18417–18425. 3

Wenqiang Lei, Xisen Jin, Min-Yen Kan, Zhaochun Ren, Xiangnan He, and Dawei Yin. 2018. Sequicity: Simplifying task-oriented dialogue systems with single sequence-to-sequence architectures. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1437–1447, Melbourne, Australia. Association for Computational Linguistics. 17

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2024a. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36. 4, 7, 8

Lei Li, Yongfeng Zhang, and Li Chen. 2021. Personalized transformer for explainable recommendation. *arXiv preprint arXiv:2105.11601*. 19

Lei Li, Yongfeng Zhang, and Li Chen. 2023a. Personalized prompt learning for explainable recommendation. *ACM Transactions on Information Systems*, 41(4):1–26. 5, 19

Pan Li, Yuyan Wang, Ed H. Chi, and Minmin Chen. 2023b. Prompt tuning large language models on personalized aspect extraction for recommendations. 5, 19

Qingyao Li, Lingyue Fu, Weiming Zhang, Xianyu Chen, Jingwei Yu, Wei Xia, Weinan Zhang, Ruiming Tang, and Yong Yu. 2023c. Adapting large language models for education: Foundational capabilities, potentials, and challenges. *arXiv preprint arXiv:2401.08664*. 8

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing. 18

Yongqi Li, Xinyu Lin, Wenjie Wang, Fuli Feng, Liang Pang, Wenjie Li, Liqiang Nie, Xiangnan He, and Tat-Seng Chua. 2024b. A survey of generative search and recommendation in the era of large language models. 5

Zekun Li, Baolin Peng, Pengcheng He, Michel Galley, Jianfeng Gao, and Xifeng Yan. 2024c. Guiding large language models via directional stimulus prompting. *Advances in Neural Information Processing Systems*, 36. 6

Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. 2024. Can large language models reason about medical questions? *Patterns*, 5(3). 3

Yen-Ting Lin and Yun-Nung Chen. 2023. LLM-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, pages 47–58, Toronto, Canada. Association for Computational Linguistics. 3

Bing Liu, Gokhan Tür, Dilek Hakkani-Tür, Pararth Shah, and Larry Heck. 2018. Dialogue learning with human teaching and feedback in end-to-end trainable task-oriented dialogue systems. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2060–2069, New Orleans, Louisiana. Association for Computational Linguistics. 17

Jerry Liu. 2022. LlamaIndex. 6

Junling Liu, Chao Liu, Peilin Zhou, Renjie Lv, Kang Zhou, and Yan Zhang. 2023. Is chatgpt a good recommender? a preliminary study. 19

Junpeng Liu, Yifan Song, Bill Yuchen Lin, Wai Lam, Graham Neubig, Yuanzhi Li, and Xiang Yue. 2024. Visualwebbench: How far have multimodal llms evolved in web page understanding and grounding? *arXiv preprint arXiv:2404.05955*. 18

Qian Liu, Yihong Chen, Bei Chen, Jian-Guang Lou, Zixuan Chen, Bin Zhou, and Dongmei Zhang. 2020. You impress me: Dialogue generation via mutual persona perception. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1417–1427, Online. Association for Computational Linguistics. 6

Ehsan Lotfi, Maxime De Bruyn, Jeska Buhmann, and Walter Daelemans. 2024. Personalitychat: Conversation distillation for personalized dialog modeling with facts and traits. 18

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2017. Hierarchical question-image co-attention for visual question answering. 6

Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. 2023. Analyzing leakage of personally identifiable information in language models. 8

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2024. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36. 1

Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. 2016. Key-value memory networks for directly reading documents. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1400–1409, Austin, Texas. Association for Computational Linguistics. 17

Madison Milne-Ives, Caroline de Cock, Ernest Lim, Melissa Harper Shehadeh, Nick de Pennington, Guy Mole, Eduardo Normando, and Edward Meinert. 2020. The effectiveness of artificial intelligence conversational agents in health care: systematic review. *Journal of medical Internet research*, 22(10):e20346. 6

Johannes E. M. Mosig, Shikib Mehri, and Thomas Kober. 2020. Star: A schema-guided dialog dataset for transfer learning. 18

Nikola Mrkšić, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 2017. Neural belief tracker: Data-driven dialogue state tracking. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1777–1788, Vancouver, Canada. Association for Computational Linguistics. 17

Nikola Mrkšić, Diarmuid O Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 2016. Neural belief tracker: Data-driven dialogue state tracking. *arXiv preprint arXiv:1606.03777*. 17

Nikola Mrkšić and Ivan Vulić. 2018. Fully statistical neural belief tracking. 17

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 188–197. 19

OpenAI. 2022. Introducing chatgpt. `https://openai.com/index/chatgpt/`. 1

Keyu Pan and Yawen Zeng. 2023. Do llms possess a personality? making the mbti test an amazing evaluation for large language models. 7

PapersWithCode. 2020. Baidu personachat dataset. `https://paperswithcode.com/dataset/baidu-personachat`. 18

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22. 2, 3

Minju Park, Sojung Kim, Seunghyun Lee, Soonwoo Kwon, and Kyuseok Kim. 2024. Empowering personalized learning through a conversation-based tutoring system with student modeling. *arXiv preprint arXiv:2403.14071*. 6

Kai Petersen, Claes Wohlin, and Dejan Baca. 2009. The waterfall model in large-scale development. In *Product-Focused Software Process Improvement: 10th International Conference, PROFES 2009, Oulu, Finland, June 15-17, 2009. Proceedings 10*, pages 386–400. Springer. 2

Chen Qian, Xin Cong, Cheng Yang, Weize Chen, Yusheng Su, Juyuan Xu, Zhiyuan Liu, and Maosong Sun. 2023. Communicative agents for software development. *arXiv preprint arXiv:2307.07924*. 2, 3, 4

Osman Ramadan, Paweł Budzianowski, and Milica Gašić. 2018. Large-scale multi-domain belief tracking with knowledge sharing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 432–437, Melbourne, Australia. Association for Computational Linguistics. 18

Abhinav Rastogi, Dilek Hakkani-Tur, and Larry Heck. 2018. Scalable multi-domain dialogue state tracking. 17

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8689–8696. 18

Chris Richardson, Yao Zhang, Kellen Gillespie, Sudipta Kar, Arshdeep Singh, Zeynab Raeesy, Omar Zia Khan, and Abhinav Sethy. 2023. Integrating summarization and retrieval for enhanced personalization via large language models. 7

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, et al. 2024. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*. 6

Alireza Salemi and Hamed Zamani. 2024. Towards a search engine for machines: Unified ranking for multiple retrieval-augmented large language models. *arXiv preprint arXiv:2405.00175*. 5

Rusheb Shah, Quentin Feuillade-Montixi, Soroush Pour, Arush Tagade, Stephen Casper, and Javier Rando. 2023. Scalable and transferable black-box jailbreaks for language models via persona modulation. 8

Nikhil Sharma, Q Vera Liao, and Ziang Xiao. 2024. Generative echo chamber? effects of llm-powered search systems on diverse information seeking. *arXiv preprint arXiv:2402.05880*. 5

Shady Shehata, David Santandreu Calonge, Philip Purnell, and Mark Thompson. 2023. Enhancing video-based learning using knowledge tracing: Personalizing students' learning experience with ORBITS. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 100–107, Toronto, Canada. Association for Computational Linguistics. 6

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2024. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36. 1, 2

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180. 6

Haoyu Song, Wei-Nan Zhang, Jingwen Hu, and Ting Liu. 2020. Generating persona consistent dialogues by exploiting natural language inference. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8878–8885. 6

Xiaoyang Song, Yuta Adachi, Jessie Feng, Mouwei Lin, Linhao Yu, Frank Li, Akshat Gupta, Gopala Anumanchipalli, and Simerjot Kaur. 2024. Identifying multiple personalities in large language models with external evaluation. 7, 8

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 196–205, Denver, Colorado. Association for Computational Linguistics. 17

Aleksandra Sorokovikova, Sharwin Rezagholi, Natalia Fedorova, and Ivan Yamshchikov. 2024. LLMs simulate big5 personality traits: Further evidence. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, pages 83–87, St. Julians, Malta. Association for Computational Linguistics. 7

Sofia Eleni Spatharioti, David M. Rothschild, Daniel G. Goldstein, and Jake M. Hofman. 2023. Comparing traditional and llm-based search for consumer choice: A randomized experiment. 5

Elizabeth C Stade, Shannon Wiltsey Stirman, Lyle H Ungar, Cody L Boland, H Andrew Schwartz, David B Yaden, João Sedoc, Robert J DeRubeis, Robb Willer, and Johannes C Eichstaedt. 2024. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *npj Mental Health Research*, 3(1):12. 6

Hiroaki Sugiyama, Masahiro Mizukami, Tsunehiro Arimoto, Hiromi Narimatsu, Yuya Chiba, Hideharu Nakajima, and Toyomi Meguro. 2021. Empirical analysis of training strategies of transformer-based japanese chit-chat systems. 18

Chenkai Sun, Ke Yang, Revanth Gangi Reddy, Yi R. Fung, Hou Pong Chan, ChengXiang Zhai, and Heng Ji. 2024. Persona-db: Efficient large language model personalization for response prediction with collaborative data refinement. 7

Weiwei Sun, Shuo Zhang, Krisztian Balog, Zhaochun Ren, Pengjie Ren, Zhumin Chen, and Maarten de Rijke. 2021. Simulating user satisfaction for the evaluation of task-oriented dialogue systems. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2499–2506. 17

M Swift and J Allen. 2010. Towards a personal health management assistant. *Journal of biomedical informatics*, 43(5):S13–S16. 8

Zhaoxuan Tan, Qingkai Zeng, Yijun Tian, Zheyuan Liu, Bing Yin, and Meng Jiang. 2024. Democratizing large language models via personalized parameter-efficient fine-tuning. 7

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2023a. Medagents: Large language models as collaborators for zero-shot medical reasoning. *arXiv preprint arXiv:2311.10537*. 3, 4

Yihong Tang, Bo Wang, Miao Fang, Dongming Zhao, Kun Huang, Ruifang He, and Yuexian Hou. 2023b. Enhancing personalized dialogue generation with contrastive latent variables: Combining sparse and dense persona. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5456–5468, Toronto, Canada. Association for Computational Linguistics. 6

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*. 17

Quan Tu, Chuanqi Chen, Jinpeng Li, Yanran Li, Shuo Shang, Dongyan Zhao, Ran Wang, and Rui Yan. 2023. Characterchat: Learning towards conversational ai with personalized social support. *arXiv preprint arXiv:2308.10278*. 18

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023a. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*. 3, 4

Jeffrey G. Wang, Jason Wang, Marvin Li, and Seth Neel. 2024a. Pandora's white-box: Increased training data leakage in open llms. *ArXiv*, abs/2402.17012. 8

Jianhong Wang, Yuan Zhang, Tae-Kyun Kim, and Yunjie Gu. 2020. Modelling hierarchical structure between dialogue policy and natural language generator with option framework for task-oriented dialogue system. *arXiv preprint arXiv:2006.06814*. 17

Lei Wang and Ee-Peng Lim. 2023. Zero-shot next-item recommendation using large pretrained language models. 5, 19

Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S Yu, and Qingsong Wen. 2024b. Large language models for education: A survey and outlook. *arXiv preprint arXiv:2403.18105*. 6

Xintao Wang, Yunze Xiao, Jen tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. 2024c. Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. 7

Yancheng Wang, Ziyan Jiang, Zheng Chen, Fan Yang, Yingxue Zhou, Eunah Cho, Xing Fan, Xiaojiang Huang, Yanbin Lu, and Yingzhen Yang. 2024d. Recmind: Large language model powered agent for recommendation. 5

Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023b. Aligning large language models with human: A survey. *arXiv preprint arXiv:2307.12966*. 3

Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2023c. Unleashing cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. *arXiv preprint arXiv:2307.05300*, 1(2):3. 7

Zhilin Wang, Yu Ying Chiu, and Yu Cheung Chiu. 2023d. Humanoid agents: Platform for simulating human-like generative agents. *arXiv preprint arXiv:2310.05418*. 3

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. 1

Wei Wei, Quoc Le, Andrew Dai, and Jia Li. 2018. AirDialogue: An environment for goal-oriented dialogue research. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3844–3854, Brussels, Belgium. Association for Computational Linguistics. 18

Cheng-Kuang Wu, Wei-Lin Chen, and Hsin-Hsi Chen. 2023a. Large language models perform diagnostic reasoning. *arXiv preprint arXiv:2307.08922*. 3

Chien-Sheng Wu, Andrea Madotto, Ehsan Hosseini-Asl, Caiming Xiong, Richard Socher, and Pascale Fung. 2019. Transferable multi-domain state generator for task-oriented dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 808–819, Florence, Italy. Association for Computational Linguistics. 17

Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, and Ming Zhou. 2020. MIND: A large-scale dataset for news recommendation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3597–3606, Online. Association for Computational Linguistics. 19

Ledell Wu, Adam Fisch, Sumit Chopra, Keith Adams, Antoine Bordes, and Jason Weston. 2018. Starspace: Embed all the things! In *Proceedings of the AAAI conference on artificial intelligence*, volume 32. 17

Ning Wu, Ming Gong, Linjun Shou, Shining Liang, and Daxin Jiang. 2023b. Large language models are diverse role-players for summarization evaluation. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 695–707. Springer. 4

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. 2023. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*. 4

Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. 2023. Baize: An open-source chat model with

parameter-efficient tuning on self-chat data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6268–6278, Singapore. Association for Computational Linguistics. 6

Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K Dey, and Dakuo Wang. 2024. Mental-llm: Leveraging large language models for mental health prediction via online text data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1):1–32. 6

Dongjie Yang, Ruifeng Yuan, Yuantao Fan, Yifei Yang, Zili Wang, Shusen Wang, and Hai Zhao. 2023a. RefGPT: Dialogue generation of GPT, by GPT, and for GPT. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2511–2535, Singapore. Association for Computational Linguistics. 6

Fan Yang, Zheng Chen, Ziyan Jiang, Eunah Cho, Xiaojiang Huang, and Yanbin Lu. 2023b. Palr: Personalization aware llms for recommendation. 5, 19

Min Yang, Weiyi Huang, Wenting Tu, Qiang Qu, Ying Shen, and Kai Lei. 2021. Multitask learning and reinforcement learning for personalized dialog generation: An empirical study. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):49–62. 17

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. 2022a. Webshop: Towards scalable real-world web interaction with grounded language agents. *Advances in Neural Information Processing Systems*, 35:20744–20757. 18

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36. 2

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022b. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*. 2

Yelp. 2013. Yelp dataset. `https://www.yelp.com/dataset`. 19

Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. 2020. MultiWOZ 2.2 : A dialogue dataset with additional annotation corrections and state tracking baselines. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 109–117, Online. Association for Computational Linguistics. 18, 20

Jianguo Zhang, Kazuma Hashimoto, Chien-Sheng Wu, Yao Wang, Philip Yu, Richard Socher, and Caiming Xiong. 2020. Find or classify? dual strategy for slot-value predictions on multi-domain dialog state tracking. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 154–167, Barcelona, Spain (Online). Association for Computational Linguistics. 17

Junjie Zhang, Ruobing Xie, Yupeng Hou, Wayne Xin Zhao, Leyu Lin, and Ji-Rong Wen. 2023. Recommendation as instruction following: A large language model empowered recommendation approach. 5, 19

Kai Zhang, Yangyang Kang, Fubang Zhao, and Xiaozhong Liu. 2024a. Llm-based medical assistant personalization with short- and long-term memory coordination. 6

Kai Zhang, Lizhi Qing, Yangyang Kang, and Xiaozhong Liu. 2024b. Personalized llm response generation with parameterized memory injection. 7

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018a. Personalizing dialogue agents: I have a dog, do you have pets too? 18

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018b. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics. 20

Boyuan Zheng, Boyu Gou, Jihyung Kil, Huan Sun, and Yu Su. 2024. Gpt-4v (ision) is a generalist web agent, if grounded. *arXiv preprint arXiv:2401.01614*. 17

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. 4

Peixiang Zhong, Chen Zhang, Hao Wang, Yong Liu, and Chunyan Miao. 2020. Towards persona-based empathetic conversational models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6556–6566, Online. Association for Computational Linguistics. 6

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. Memorybank: Enhancing large language models with long-term memory. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19724–19731. 7

Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Yonatan Bisk, Daniel Fried, Uri Alon, et al. 2023. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*. 18

Yujia Zhou, Zhicheng Dou, Bingzheng Wei, and Ruobing Xievand Ji-Rong Wen. 2021. Group based personalized search by integrating search behaviour and friend network. 5

Yujia Zhou, Qiannan Zhu, Jiajie Jin, and Zhicheng Dou. 2024. Cognitive personalized search integrating large language models with an efficient memory mechanism. *arXiv preprint arXiv:2402.10548*. 5

Noah Ziems, Wenhao Yu, Zhihan Zhang, and Meng Jiang. 2023. Large language models are built-in autoregressive search engines. 5

A Web

In this environment, LLMs operate web navigation autonomously, performing actions such as clicking items, capturing contents, and searching from external knowledge on the web, without a specific persona assigned. Certainly, web tasks involve two key components: *HTML understanding* and *visual grounding*, which are highly related to the effectiveness of web agents (Zheng et al., 2024; Koh et al., 2024). Meanwhile, a stream of works, compiled in Table 1, proposes several benchmarks to assess web agents in diverse aspects.

HTML Understanding. Kim et al. (2024a) showcase that the ability of HTML understanding is inherent in LLMs with the Recursive Criticism and Improvement (RCI) prompting method. However, due to the special formats and long context elements of HTML which are hard for LLMs to process and respond accurately, most research enhances this capability through fine-tuning methods (Gur et al., 2022, 2023; Deng et al., 2024).

Visual Grounding. Another line of research focuses on the visual grounding aspect of HTML understanding, which directly operates on rendered webpages instead of the HTML source code. Some literature proposes web agent frameworks, such as CogAgent (Hong et al., 2023b) and SeeClick (Cheng et al., 2024), leveraging Large Multi-modal Models (LMMs) (Achiam et al., 2023; Team et al., 2023). With additional information from webpage screenshots, LMMs usually outperform text-based LLMs (Zheng et al., 2024).

B Background of Dialogue Generation

ToD modeling typically consists of four modules: Natural Language Understanding (NLU), Dialogue State Tracker (DST), Dialogue Policy (DP), and Natural Language Generator (NLG). Among these, DST (Jacqmin et al., 2022) plays a crucial role in modeling a multi-turn dialogue while updating the state of the conversation. The state could include user information, preferences, and goals. Mrkšić et al. (2016) propose a novel method known as Neural Belief Tracker (NBT), which features an enhanced version of update mechanisms as described in Mrkšić and Vulić (2018). This method advances representation learning by predicting and updating various aspects of the user's requests and goals through *belief tracking*.

Prior to the rise of LLMs, many models focus on improving different parts of the module, such as state tracking (Mrkšić et al., 2017; Rastogi et al., 2018; Wu et al., 2019; Zhang et al., 2020), while others concentrate on policy optimization using ground-truth dialogue states (Wang et al., 2020; Sun et al., 2021). Various attempts tried to combine different modules to create a fully end-to-end ToD modeling: Liu et al. (2018); Yang et al. (2021) use reinforcement learning (RL) to combine DP and NLG. Lei et al. (2018) combine DST and NLG with a sequence-to-sequence approach. Huang et al. (2020) propose a method based on the variational autoencoder (VQ-VAE) framework (Kingma and Welling, 2013) and use three-stage learning, including Semantic Latent Action Learning, Action Alignment across Domains, and Domain-Specific Action Learning. Finally, the SIMPLETOD (Hosseini-Asl et al., 2020) model integrates different sub-tasks in a unified end-to-end manner, paving the way for fully LLM-based approaches in ToD modeling.

On the other hand, in the pre-LLM era, User persona modeling used methods like Sordoni et al. (2015) and STARSPACE (Wu et al., 2018) to rank the most similar utterance in the dataset and generate a candidate reply. Additionally, Miller et al. (2016) enhanced its ability by considering the dialogue history.

Benchmark	#Instances	#Domains	Realistic Env.	Dynamic Interaction	Visual Needed	Assessment
WebShop (Yao et al., 2022a)	12,087	1	✗	✓	✗	End-to-end
Mind2Web (Deng et al., 2024)	2,350	5	✓	✗	✗	End-to-end
WebArena (Zhou et al., 2023)	812	4	✓	✓	✗	End-to-end
VisualWebArena (Koh et al., 2024)	910	3	✓	✓	✓	End-to-end
VisualWebBench (Liu et al., 2024)	1,500	12	✓	✗	✓	Fine-grained

Table 1: Comparison between recent benchmarks in the web environment. *Realistic Env.* denotes whether the benchmark's environments are based on actual web pages or realistic web navigation simulations. *Dynamic Interaction* indicates whether the benchmark supports dynamic interactions rather than remaining in static states. *Visual Needed* denotes whether the benchmark involves visually grounded tasks. *Assessment* refers to the types of assessment. An end-to-end benchmark includes tasks with simple instructions, requiring step-by-step solutions to reach the final answers. A fine-grained benchmark contains tasks with a detailed assessment of essential skills in the web environment such as Optical Character Recognition (OCR), and semantic understanding.

Category	Dataset	#Dialogues	#Utterance	#Domains
ToD	MultiWOZ 1.0 (Budzianowski et al., 2018)	10,438	75,894	7
	MultiWOZ 2.0 (Ramadan et al., 2018)	8,438	63,841	7
	MultiWOZ 2.1 (Eric et al., 2020)	7,032	57,022	7
	MultiWOZ 2.2 (Zang et al., 2020)	10,438	71,572	7
	SGD (Rastogi et al., 2020)	22,825	463,284	20
	STAR (Mosig et al., 2020)	6,652	127,833	13
	AirDialogue (Wei et al., 2018)	4,000	52,000	1
	UniDA (He et al., 2022)	70,726	975,780	13
User Persona	PersonaChat (Zhang et al., 2018a)	11,907	164,356	1
	ConvAI2 (Dinan et al., 2019)	13,500	182,150	1
	Baidu PersonaChat (PapersWithCode, 2020)	20,000	280,000	1
	JPersonaChat (Sugiyama et al., 2021)	10,000	140,000	1
	JEmpatheticDialogues (Sugiyama et al., 2021)	25,000	350,000	1
	DailyDialog (Li et al., 2017)	13,118	102,979	10

Table 2: A list of commonly used datasets for ToD modeling and user persona modeling. Among them, different versions of MultiWOZ (Budzianowski et al., 2018; Ramadan et al., 2018; Eric et al., 2020; Zang et al., 2020) and PersonaChat (Zhang et al., 2018a) are the most commonly used. Updated versions of MultiWOZ improve in several aspects: data quality, dialogue complexity, schema and ontology updates, and dataset sizes. PersonaChat contains various persona profiles, consisting of background, preferences, and personality traits. These profiles enable the modeling of coherent and contextual multi-turn diverse dialogue scenarios. For applications in user persona modeling, Tu et al. (2023) match individuals with persona-compatible virtual supporters and introduces the MBTI-S2Conv dataset, containing conversations between characters with distinct profiles. Lotfi et al. (2024) and Han et al. (2024) both propose synthetic datasets related to the Big Five personality.

Paper	Scene	Dataset	Method	Task
Li et al. (2023b)	Hotel, Movies & TV, Restaurant	TripAdvisor, Amazon, Yelp	Embeddings, Prompting, Fine-tuning	Aspect extraction, Rating Prediction
P5 (Geng et al., 2022)	Sports, Beauty, Toys, Yelp	Amazon (Ni et al., 2019), Yelp	Pretraining, Prompting	Rating Prediction, Sequential Recommendation, Explanation Generation, Review Generation, and Direct Recommendation
PETER Li et al. (2021)	Hotel, Movies & TV, Restaurant	TripAdvisor, Amazon, Yelp	Transformer	Rating prediction and Explanation Generation
PEPLER (Li et al., 2023a)	Hotel, Movies, TV and Restaurant	TripAdvisor5 (Hotel), Amazon (movies& TV) and Yelp7 (restaurant)	Prompting, Fine-tuning	Explanation Generation
PALR (Yang et al., 2023b)	Movies, Beauty	MovieLens-1M (Harper and Konstan, 2015), Amazon Beauty (Ni et al., 2019)	Fine-tuning, User Profile Generation, Retrieval	User Profile Generation and Direct Recommendation
Chu et al. (2023)	Sports, Outdoors, Beauty, Toys and Games	Amazon	Fine-tuning	Rating Prediction, Sequential Recommendation, Direct Recommendation, Explanation Generation and Review Summarization
Liu et al. (2023)	Beauty	Amazon	Prompting	Rating Prediction, Sequential Recommendation, Direct Recommendation, Explanation Generation and Review Summarization
Zhang et al. (2023)	Video Games	Amazon	Instruction tuning	Sequential Recommendation and Direct Recommendation
Hou et al. (2024)	Movies	Amazon (Ni et al., 2019), MovieLens-1M Harper and Konstan (2015)	Prompting	Sequential Recommendation
Wang and Lim (2023)	Movies	MovieLens-1M (Harper and Konstan, 2015)	Prompting	Sequential Recommendation and Direct Recommendation
Chen et al. (2022)	News	MIND (Wu et al., 2020), Reddit	Fine-tuning with weak labels	Direct Recommendation

Table 3: An overview of existing research in recommendation. Following the classification of Liu et al. (2023), we classify recommendation systems into five types: rating prediction, sequential recommendation, explanation Generation, and review generation, and direct recommendation.

Dataset	Scene	Task	#Instances	#Users	#Items
Amazon Review (Ni et al., 2019)	Products	Ratings, Reviews	233.1M	43.53M	15.17M
MovieLens (Harper and Konstan, 2015)	Movies	Ratings	100,000	1,000	1,700
Yelp (Yelp, 2013)	Businesses	Ratings & Reviews	6,990,280	1,987,897	150,346
TripAdvisor (Li et al., 2023a)	Hotels, Restaurants	Ratings & Reviews	320,023	9,765	6,280
MIND (Wu et al., 2020)	News	Sequence recommendation	15M	1M	160k

Table 4: A list of commonly used datasets in personalized LLMs for recommendation and search task. For the fifth column, the instances include reviews and ratings.

Domain	Dialogue, Slots, and State
Restaurant	[USER:] I need a place to dine in the center that's expensive. **Slots:** restaurant-area: centre, restaurant-pricerange: expensive **State:** active_intent: find_restaurant
	[SYSTEM:] I have several options for you; do you prefer African, Asian, or British food? **State:** active_intent: find_restaurant, requested_slots: restaurant-food
	[USER:] Any sort of food would be fine, as long as it is a bit expensive. Could I get the phone number for your recommendation? **Slots:** restaurant-area: centre, restaurant-pricerange: expensive, restaurant-food **State:** active_intent: find_restaurant, requested_slots: restaurant-phone
	[SYSTEM:] There is an African place named Bedouin in the center. How does that sound? **Slots:** restaurant-name: Bedouin
	[USER:] Sounds good, could I get that phone number? Also, could you recommend me an expensive hotel? **Slots:** restaurant-area: centre, restaurant-name: Bedouin, restaurant-pricerange: expensive, restaurant-phone **State:** active_intent: find_restaurant, requested_slots: restaurant-phone
Hotel	[SYSTEM:] Bedouin's phone is 01223367660. As far as hotels go, I recommend the University Arms Hotel in the center of town. **Slots:** hotel-name: University Arms Hotel **State:** active_intent: find_hotel
	[USER:] Yes. Can you book it for me? **State:** active_intent: find_hotel, requested_slots: hotel-booking
	[SYSTEM:] Sure, when would you like that reservation?
	[USER:] I want to book it for 2 people and 2 nights starting from Saturday. **Slots:** hotel-bookday: Saturday, hotel-bookpeople: 2, hotel-bookstay: 2 **State:** active_intent: book_hotel, slot_values: hotel-bookday: Saturday, hotel-bookpeople: 2, hotel-bookstay: 2
	[SYSTEM:] Your booking was successful. Your reference number is FRGZWQL2. May I help you further? **Slots:** hotel-bookday: Saturday, hotel-bookpeople: 2, hotel-bookstay: 2, reference number: FRGZWQL2
	[USER:] That is all I need to know. Thanks, goodbye. [SYSTEM:] Thank you so much for Cambridge TownInfo center. Have a great day!

Table 5: An example of ToD modeling from the MultiWOZ dataset (Zang et al., 2020).

Persona	Chat
I fly airplanes. I enjoy building computers. My favorite band is tool. I am in the army. I dropped out of college.	[PERSON 1:] Hello how are u tonight [PERSON 2:] Hi. I am okay. tired, but okay. how are you ? [PERSON 1:] I am doing good should be sleeping i have school but can't sleep [PERSON 2:] I did not finish school, I enlisted in the army instead . [PERSON 1:] Wow I am only 14 so I can't do that just yet but I hope too [PERSON 2:] nice. stay in school and work hard . [PERSON 1:] I try i like video games and race cars [PERSON 2:] I like video games too, fallout is my favorite. [PERSON 1:] I am a call of duty girl i can't wait for the new one [PERSON 2:] My younger brother is a cod player too. he is pretty good . [PERSON 1:] I have three best friends but lots of other friends that play it [PERSON 2:] I have a best friend, she is a pilot like me. [PERSON 1:] What kind of plane do u fly [PERSON 2:] A bomber, it is awesome. do you want to take lessons [PERSON 1:] I am kinda afraid of heights so not sure flying is for me [PERSON 2:] You should at least try to go up in a plane, it is a blast.

Table 6: An example of user persona modeling (§3.5) from Persona-Chat dataset (Zhang et al., 2018b).