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

Yu-Min Tseng[*αβ] **Yu-Chao Huang**[*α] **Teng-Yun Hsiao**[*α]
Yu-Ching Hsu[α] **Jia-Yin Foo**[α] **Chao-Wei Huang**[α] **Yun-Nung Chen**[α]

[α]National Taiwan University [β]Academia Sinica
ymtseng@nlg.csie.ntu.edu.tw
{R11222015, B10502058}@ntu.edu.tw
y.v.chen@ieee.org

Abstract

Recently, methods investigating how to adapt large language models (LLMs) for specific scenarios have gained great attention. Particularly, the concept of *persona*, originally adopted in dialogue literature, has re-surged as a promising avenue. However, the growing research on persona is relatively disorganized, lacking a systematic overview. 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. To the best of our knowledge, we present the first survey tailored for LLM role-playing and LLM personalization under the unified view of persona, including taxonomy, current challenges, and potential directions. To foster future endeavors, we actively maintain a paper collection available to the community.[1]

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 scenarios* 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). With personas, LLMs are equipped to generate tailored responses while harnessing the full capabilities. However, the growing literature on persona in the LLM era is relatively disorganized, lacking a unifying overview.

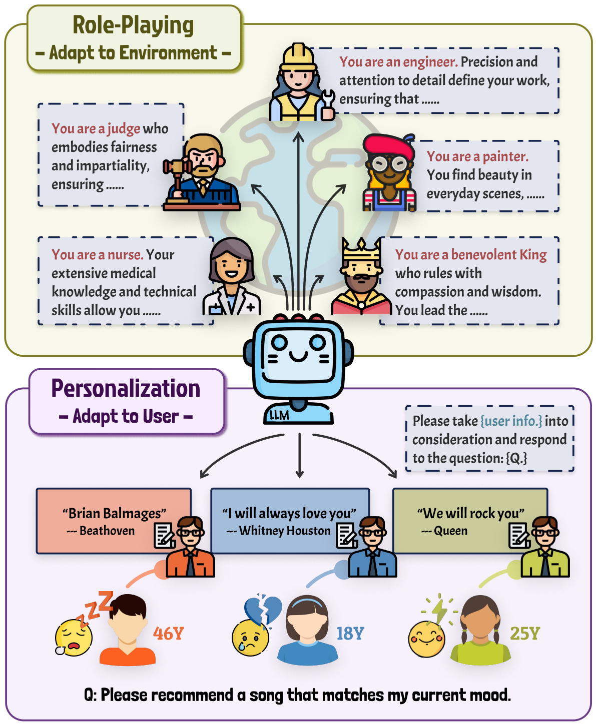

Figure 1: An illustrated overview of the survey. In *Role-Playing*, LLMs act according to assigned personas (*i.e.*, roles) under a defined environment. For example, given role names (*red*) with descriptions (*black*), LLMs are tasked to role-play in a social simulation scenario. For *Personalization*, LLMs consider users' personas (*blue*) to generate tailored responses to the same question. The dashed rectangles indicate prompts and the solid rectangles indicate LLMs' responses.

In this paper, we aim to close the gap by offering a comprehensive survey and a systematic categorization of present studies. Specifically, we divide current research into two main directions, 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. Detailed definitions and key design aspects of each are:

[1]https://github.com/MiuLab/PersonaLLM-Survey
[*]Equal contribution.

- ***LLM Role-Playing.*** LLMs are tasked to play the assigned personas (*i.e.,* roles) and act accordance to environmental feedback. The key aspect is *how LLMs adapt to defined environments*.

- ***LLM Personalization.*** LLMs are tasked to take care of users' personas (*e.g.*, background information, or historical behaviors) to meet customized needs. The key aspect is *how LLMs adapt to distinct users*.

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 hope this work could serve as a beginner guide for newcomers and a practical roadmap for current researchers in the field.

Overview. We organize the survey paper as laid out in the taxonomy (Figure 2). Firstly, we introduce LLM role-playing (§2), including agent-based role-playing of LLMs (§2.1), various environments LLMs adapt to (§2.2), and how LLMs interact within the environments as well as emergent behaviors stemmed from their interactions (§2.3). Subsequently, we elaborate LLM personalization (§3), outlining the diverse tasks (§3.1) and methods (§3.2) for personalizing LLMs. The benchmarks and datasets for various environments and personalized scenarios are introduced in relevant sections accordingly. We also add a separate branch (§4) for LLM personality evaluation focusing on psychological assessments. Lastly, we highlight current challenges in role-playing and personalized LLMs and suggest potential future directions (§5).

2 LLM Role-Playing

In order to perform diverse tasks effectively, a potential solution is to train specialized LLMs tailored to each task. However, this process is computationally prohibitive and time-consuming. To address this, recent studies have demonstrated that leveraging *role-playing* increases LLMs capabilities, including reasoning (Wu et al., 2023a; Kong et al., 2023), planning (Huang et al., 2024; Li et al., 2024b), and problem-solving (Wang et al., 2023e; Qian et al., 2023). Specifically, role-playing triggers the corresponding inherent personalities of LLMs to generate responses aligned with given roles and environments. Equipped with assigned personas (*i.e.*, roles), LLMs adapt adequately to defined environments, thereby enhancing their performance.

In this section, we first clarify the relationship between LLM role-playing and LLM-based agent (§2.1). Next, we introduce diverse environments that LLMs could adapt to (§2.2). Lastly, we identify agentic interactions between LLMs within defined environments and emergent behaviors stemming from their interactions (§2.3).

2.1 LLM-based Agent

Recently, several research has explored the planning, reflection, and tool-using abilities in LLMs, potentially paving the way for artificial general intelligence (AGI; Bubeck et al., 2023). Thus, *LLM-based agent* has surged as an fast-growing research field (Wang et al., 2024b; Xi et al., 2023). Moreover, the *Multi-Agent* framework, where multiple LLM-based agents cooperate and communicate with each other in order to solve complex tasks, has also appeared as a prevalent research direction (Guo et al., 2024).

To answer the key question discussed in (§1): *how LLMs adapt to defined environments*, the predominant approach is LLM-based agents with role-playing, where LLMs act according to the assigned personas to interact with the environments. For instance, one of the first works of LLM-based agents proposes *generative agents* (Park et al., 2023), which adopts role-playing to mimic human behaviors. By giving a list of sentences with name, age, and personality trait that specifies the role, generative agents can perform accordingly and engage in a social simulation environment.

2.2 Environments

One key aspect of the agentic frameworks is the environments, within which the LLMs could operate and interact given certain rules and receive feedback. We classify the environments for LLM role-playing into five categories: software development, web, game, medical application, and LLM as evaluators, detailed in (§2.2.1) to (§2.2.5). Furthermore, we explore a range of frameworks that generalize across different environments in (§2.2.6). Figure 3 demonstrates an illustration by providing a simple example task of each environment.

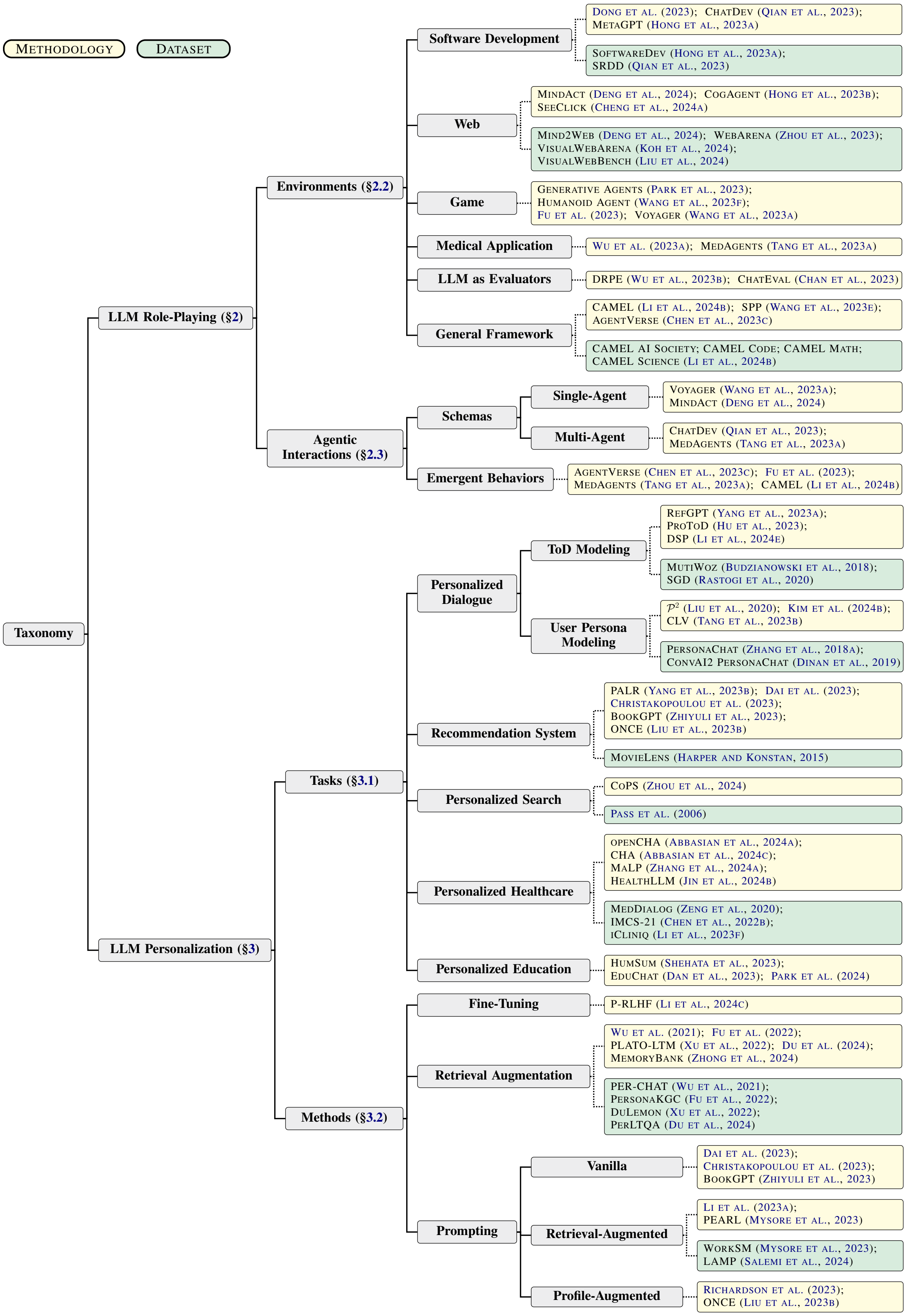

Figure 2: The taxonomy of LLM role-playing and LLM personalization.

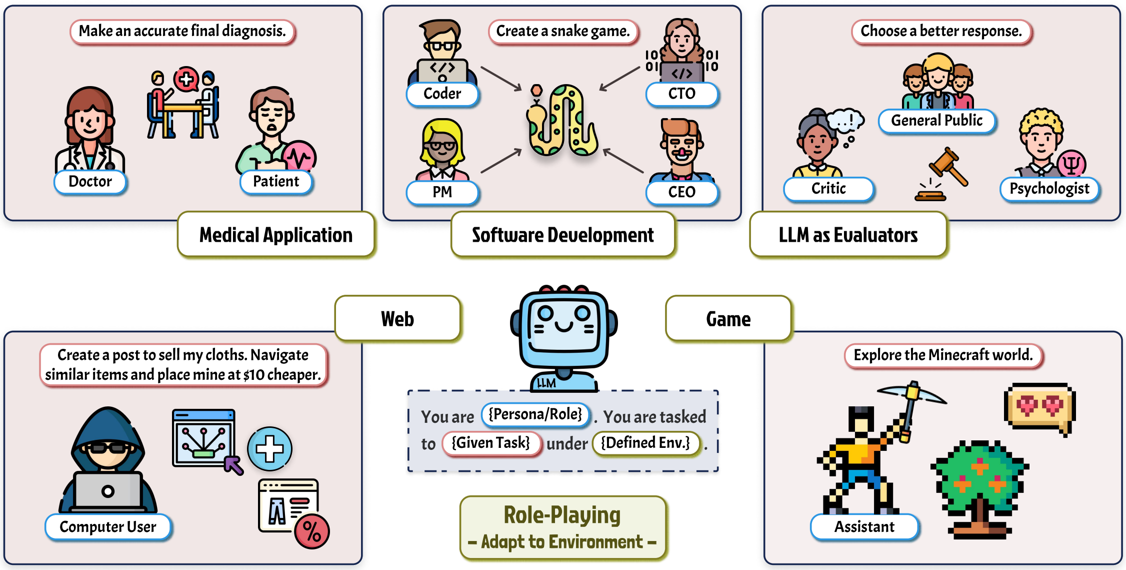

Figure 3: An illustration of five LLM role-playing environments: *Software Development* (§2.2.1), *Web* (§2.2.2), *Game* (§2.2.3), *Medical Application* (§2.2.4), and *LLM as Evaluators* (§2.2.5). For each environment, we provide a simple scenario with a task description (*red-bordered*) and relevant personas (*i.e.*, roles; *blue-bordered*). The dashed rectangle represents an example LLM role-playing prompt template. In addition to the above environments, several research also proposes general frameworks applicable to different environments (§2.2.6).

2.2.1 Software Development

The tasks of software development are to design programs, software, or coding projects. Example tasks include *"Create a snake game."*, *"Create a Python program to develop an interactive weather dashboard."* (Hong et al., 2023a), and HumanEval (Chen et al., 2021). Due to the complexity of these tasks, often too intricate to be completed correctly on the first attempt, existing research predominantly leverages 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 software development projects, a collaborative multi-agent framework with LLM role-playing operating as a company serves as an effective approach for software development environment (Qian et al., 2023; Hong et al., 2023a; Dong et al., 2023). Concretely, the roles might include Chief Technology Officer (CTO), Chief Product Officer (CPO), Chief Executive Officer (CEO), Product Manager, Engineer, Reviewer, and Tester. By assigning specific roles, LLMs are capable of generating content in a step-by-step and accurate manner.

A recent work by Dong et al. (2023) proposes 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 works, MetaGPT (Hong et al., 2023a) requires LLM agents to generate structured outputs instead of free-text, which demonstrates a significant increase in the success rate of target code generation.

2.2.2 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.

Benchmark	#Instances	#Domains	Realistic Env.	Dynamic Interaction	Visual Needed	Assessment
WebShop (Yao et al., 2022)	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.

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., 2024a), 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).

2.2.3 Game

Prior to the rise of LLMs, foundation works in agent-based game playing are AlphaGo (Silver et al., 2016) and AlphaGo Zero (Silver et al., 2017). Additionally, *Deal or No Deal?* (Lewis et al., 2017), which involved bargaining between two agents based on Recurrent Neural Network (RNN) models, also demonstrated the potential capabilities of agents in game playing. More recently, LLMs have surged as a more advanced technique for agents in a variety of game environments, including Minecraft (Wang et al., 2023a), social simulation (Park et al., 2023; Wang et al., 2023f), and bargaining game (Fu et al., 2023).

Within these environments, LLMs are able 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). Especially, gaming environments usually contain a wide range of information, including settings, utilizable tools, and nearby situations, which is hard for LLMs to memorize and response. Thus, memory stream with retrieval-based approaches are a crucial component for the effectiveness of LLM-based agents' role-play in the game environments (Park et al., 2023; Wang et al., 2023a).

2.2.4 Medical Application

Prior works have shown the potential utility of LLMs in medicine, establishing the stepping stone of unleashing the domain-specific ability of LLMs in this environment (Singhal et al., 2023; Thirunavukarasu et al., 2023). Wu et al. (2023a) propose DR-CoT as the first approach of leveraging LLM role-playing on knowledge-intensive, multi-step *diagnostic reasoning*, exhibited a striking improvement from standard prompting. The following 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.

Both studies assign medically relevant personas to LLMs, ranging from general roles like Doctor and Patient to specific ones such as Neurology Expert and Psychiatry Expert. Their research demonstrates that LLMs inherently possess medical knowledge (Liévin et al., 2024), thus enhancing the performance through LLM role-playing successfully. Nonetheless, there is still ample room for further exploration of LLM role-playing in a medical application environment.

2.2.5 LLM as Evaluators

As LLMs grow increasingly powerful, the evaluation of generated content becomes an extremely vital and challenging issue. Traditional n-gram metrics and model-based evaluations, such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and BERTScore (Zhang et al., 2019), cannot appropriately assess alignment with human intent (Wang et al., 2023b; Chang et al., 2023b). For instance, a summary with a high ROUGE score might not be a fluent and meaningful passage, potentially misrepresenting the original article.

Therefore, the concept of *LLM as evaluators* (*i.e.*, *LLM-as-a-judge*) has been introduced as an alternative way of evaluation. Specifically, LLMs are capable of assessing human-like values of model responses, such as empathy, fluency, and reasonableness. Several studies also demonstrate that the judgments made by LLMs have a higher correlation with human ground truth than traditional metrics (Chiang and Lee, 2023; Zheng et al., 2023; Wang et al., 2023d; Lin and Chen, 2023).

Aiming for greater similarity with human evaluation, roles in this environment span a broad spectrum representing various perspectives of human beings in society, such as the general public, critic, and news author. Wu et al. (2023b) propose the DRPE method to assess the quality of summarization by assigning statically objective roles, and dynamically subjective roles based on task settings to LLMs. Another work, ChatEval (Chan et al., 2023), further adds discussion rounds within roles to improve the evaluation process, simulating a judge group in reality. In this way, LLM as evaluators could serve as an alternative solution of evaluation while models become stronger.

2.2.6 General Framework

Despite the effectiveness of the aforementioned LLM role-playing frameworks tailored to each task, a line of research indicates that these personas heavily rely on human assignment, requiring prior knowledge and understanding of the tasks (Chen et al., 2023c). Consequently, enhancing the generalizability of these frameworks and employing automatic prompt crafting appear as potential solutions to these limitations (Li et al., 2024b; Wang et al., 2023e).

For example, Li et al. (2024b) propose a novel communicative agent framework that allows LLM agents to collaborate autonomously by *Inception Prompting*, requiring minimal human intervention. Subsequently, Wang et al. (2023e) introduce *Solo Performance Prompting (SPP)*, a method for LLMs to automatically identify personas based on given problems, and further engage in multi-turn self-collaboration with those personas. Another work, proposed by Chen et al. (2023c), is a multi-agent framework called AgentVerse that enables LLMs to dynamically adjust the personas of group members based on current progress. Overall, these frameworks demonstrate significant performance and generalizability across a range of tasks, including reasoning and generation.

2.3 Agentic Interactions

In this section, we first present two distinct schemas of LLM role-playing in (§2.3.1), and further discuss the emergent behaviors of LLMs in (§2.3.2).

2.3.1 Schemas

Across various environments, we categorize LLM role-playing schemas into two catogories: *Single-Agent* and *Multi-Agent*. The distinction is whether agents could achieve the given task single-handedly in the defined environments. Notably, each agent is capable of planning, reflecting, and decision-making (Wang et al., 2024b; Xi et al., 2023). Detailed definitions are as follows.

Single-Agent Schema. We define the single agent schema as: one agent is able to achieve its goal independently without assistance from others, even though it may coexist with multiple agents in the defined environment.

Under the notion of LLM role-playing, single agent schema mainly appears in game and web environments, where LLMs pay more attention on 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 Schema. In comparison, we define the multi-agent schema as one agent necessitated to collaborate, communicate, and ultimately achieve its goal with the support of other agents.

Software development, medical applications, and LLM as evaluators are primary environments for this schema. Similar to the real world, the essence of these environments is human interaction. LLMs impersonate various critical roles and their functionalities by adopting relevant personas, subsequently forming a multi-agent group to tackle given tasks within the defined environments. For example, 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.

As identified in prior surveys of LLM-based agents (Xi et al., 2023; Guo et al., 2024), there are two collaboration paradigms in the multi-agent schema: *Cooperative* and *Adversarial*. The cooperative paradigm facilitates information sharing among agents, with some frameworks using message pools to store each agent's current state and ongoing tasks (Hong et al., 2023a; Tang et al., 2023a; Chen et al., 2023c). On the other hand, the adversarial paradigm, including debate, competition, and criticism, enhances the decision-making process and seeks more advantages through adopting opposing perspectives (Chan et al., 2023; Fu et al., 2023).

2.3.2 Emergent Behaviors

Recent research has explored emergent behaviors in LLM role-playing (Chen et al., 2023c; Li et al., 2024b; Wang et al., 2023e; Tang et al., 2023a; Fu et al., 2023). Furthermore, as mentioned in Cheng et al. (2024b), LLM multi-agent collaboration can reflect human society phenomena or theories in a social psychology view, such as conformity and consensus reaching. To provide a pioneering perspective for future research, we integrate three collaborative behaviors in this section, mainly following Chen et al. (2023c) and others' findings.

Volunteer Behavior. This behavior primarily occurs in the multi-agent schema with the cooperative collaboration paradigm. To accomplish a team goal, LLMs proactively assist their peers or inquire if there is anything they can help with. Moreover, they contribute resources to each other, such as unallocated time and possessed materials. Through this volunteer behavior, LLMs enhance team efficiency and demonstrate cohesion and commitment within defined environments (Chen et al., 2023c; Hong et al., 2023a).

Conformity Behavior. This behavior occurs in situations where a LLM deviates from the team goal. After receiving criticism and suggestions from others, the deviating LLM refines and adjusts its behavior or decisions to better cooperate with the team. Through this conformity behavior, 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, LLMs may exhibit a *Bad Mind* that seeks to control the world (Li et al., 2024b). Furthermore, LLMs consistently display toxicity or reveal deep-seated stereotypical biases when equipping personas (Deshpande et al., 2023; Gupta et al., 2023). Such destructive behavior raises safety and bias concerns of LLM role-playing. Thus, mitigating these issues is also outlined in (§5) as a potential future direction.

3 LLM Personalization

A crucial aspect of developing strong and capable large language models is further shaping them to meet our personalized needs. Personalized LLMs have gained significant interest in recent years, leading to the organization of various workshops (Chen et al., 2023e; Deshpande et al., 2024) and competitions (Yusupov and Kuratov, 2018; Dinan et al., 2019). These events have introduced a wealth of innovative concepts for tailoring LLMs to individual user needs. Most LLMs use reinforcement learning from human feedback to fine-tune the model, aligning the output with human expectations and preferences. However, in the process, they infuse collective consciousness into the model because of biases in human-annotated data. To make LLMs more context-sensitive and enhance the human experience by adapting to individual consciousness, LLMs need to consider individual information, historical behavior, and user personas.

In this section, we first introduce five tasks of personalized LLMs (§3.1). Subsequently, we discuss various methods used in personalized LLMs (§3.2), including fine-tuning methods, retrieval augmentation, and prompting methods.

3.1 Tasks

We present various aspects of personalizing LLMs across diverse tasks, including personalized dialogue (§3.1.1), personalized recommendation system (§3.1.2), personalized search (§3.1.3), personalized healthcare (§3.1.4), and personalized education (§3.1.5), as illustrated in Figure 4.

3.1.1 Personalized Dialogue Generation

One of the most active areas for personalized LLMs is personalized dialogue generation, which can be categorized into two distinct goals: (1) generating personalized task-oriented responses and (2) modeling users' personas.

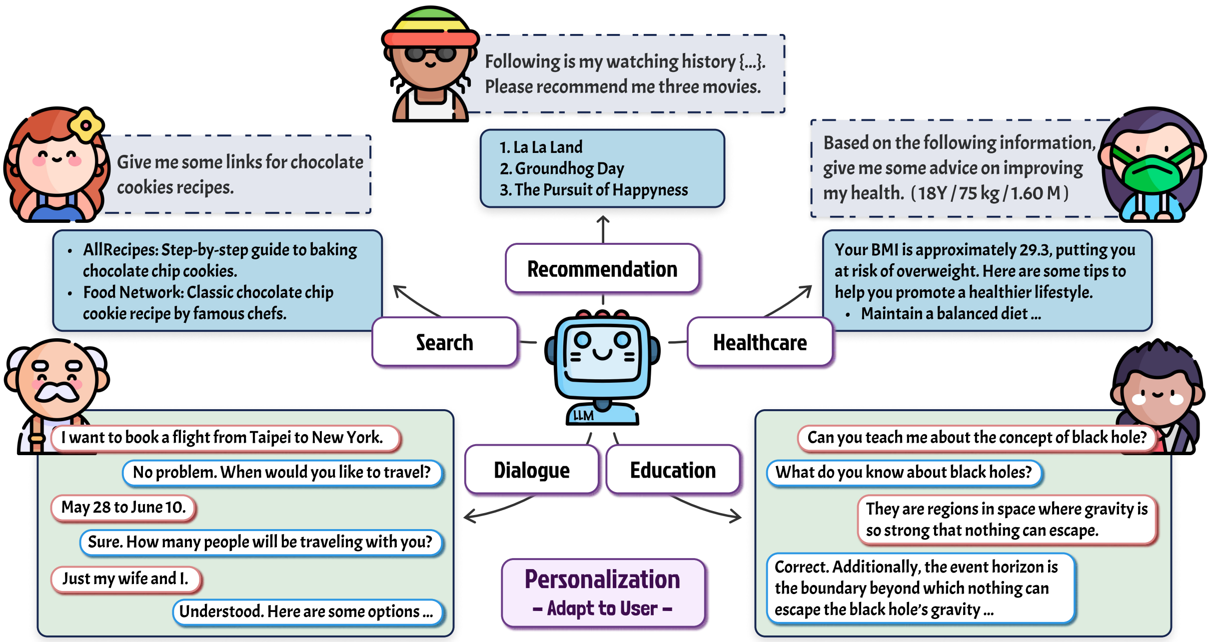

Figure 4: An illustration of five types of personalized LLMs: *Personalized Dialogue* (§3.1.1), *Personalized Search* (§3.1.3), *Recommendation System* (§3.1.2), *Personalized Healthcare* (§3.1.4), and *Personalized Education* (§3.1.5). For dialogue and education tasks, we illustrate multi-turn interactions between LLMs and users. For other tasks, the dashed rectangles represent prompts, and the solid rectangles represent the responses of LLMs.

(1) **Task-oriented Dialogue Modeling (ToD Modeling)**. ToD modeling guides users in completing specific tasks through multiple interactive steps, such as hotel bookings and restaurant reservations. It consists of understanding the semantic meaning of user utterances, following up with policy planning to determine the dialogue act, and finally generating an active response in form of natural language. See an example in Appendix A.1.

(2) **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 A.2.

Background of Personalized 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., 2019b; 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. (2021a) 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.

ToD Modeling. Madotto et al. (2020) demonstrate promising results for the few-shot ability of LLMs on each module of NLU, DST, DP, and NLG, showcasing pre-trained LLMs as a favorable end-to-end approach. Early works on creating a fully end-to-end model include SIMPLETOD (Hosseini-Asl et al., 2020) and SOLOIST (Peng et al., 2021) These models handle all sub-tasks in an end-to-end, single-sequence prediction setting on pre-trained LLMs. Similar work UBAR (Yang et al., 2021b) models the entire dialogue history, including belief states, system acts, and system responses as the input sequence. DIRECTIONAL STIMULUS PROMPTING (DSP) (Li et al., 2024e) trains a small policy model to generate hints and guide LLMs in completing tasks. In PROTOD (Hu et al., 2023), inspired by DSP (Li et al., 2024e), they propose a framework that incorporates future action hints and a policy model trained with supervised learning and reinforcement learning. Hudeček and Dusek (2023) employ in-context learning, retrieval, and state tracking with instruction fine-tuned LLM. REFGPT (Yang et al., 2023a) generates multi-turn, truthful dialogues by augmenting the dialogue history with factual and reliable sources and using prompts to guide the LLM in structuring according to predefined dialogue settings.

Some models combine reinforcement learning, focusing on the policy model (Wu et al., 2023c), the dialogue state tracker and natural language generator (Bang et al., 2023), designing reward functions for end-to-end ToD agents (Feng et al., 2023), or augmenting prompts (Li et al., 2024e; Hu et al., 2023).

Other works use LLMs to generate multi-turn dialogue datasets for training language models (Yang et al., 2023a; Huryn et al., 2022; Xu et al., 2023). On the application side, personalized dialogue has also been used in procedural content generation (PCG) (Ashby et al., 2023) in video games for customized dialogue generation.

User Persona Modeling. Zhang et al. (2018a) present one of the earlier works on personalized dialogue generation by training DIALOGPT on Reddit comments and achieving human-like performance in a single-turn dialogue setting. Later on, Zheng et al. (2019) concatenates word embedding, positional embedding, and attribute embedding and uses an encoder-decoder to pre-train a language model. Wu et al. (2019a) propose a variational autoencoder (VAE) to generate persona-aware responses. They also introduce three persona evaluation metrics: *uRank*, which measures the user's language style; *uPPL*, which imitates the language style of the user; and *uDistinct*, which measures the diversity between different users.

More sophisticated models, such as COBERT (Zhong et al., 2020), propose persona-based empathetic conversations using BERT as the encoder and a two-hop co-attention mechanism (Lu et al., 2017) to refine these embeddings and identify the most relevant response given the context and persona information. They further present a persona-based empathetic conversation (PEC) dataset. Other works take different approaches. For example, Song et al. (2020) utilizes natural language inference (NLI) as an RL task with response-persona as a reward to generate persona-consistent dialogue. Liu et al. (2020) proposes $\mathcal{P}^2$, a mutual persona perception model, and employs supervised training and self-play fine-tuning during the training process. Tang et al. (2023b) combine sparse persona descriptions, dense persona descriptions, and dialogue history to generate personalized responses.

Datasets. In Table 2, we organize some common datasets used for ToD Modeling and Persona Modeling. Among them, various versions of MultiWOZ (Budzianowski et al., 2018; Ramadan et al., 2018; Eric et al., 2020; Zang et al., 2020) and PersonaChat (Zhang et al., 2018b) 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.

On the potential application side for persona modeling, Tu et al. (2023) matches individuals with persona-compatible virtual supporters and introduces the MBTI-S2Conv dataset, which contains conversations between characters with distinct profiles. On the side of creating synthetic persona-related datasets, Lotfi et al. (2024) distill a dataset based on PersonaChat (Zhang et al., 2018b) with

Category	Dataset	#Dialogues	#Utterance	#Domains
ToD (§3.1.1)	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
Persona (§3.1.1)	PersonaChat (Zhang et al., 2018b)	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 and Persona Modeling.

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., 2023c)	Hotels, Restaurants	Ratings & Reviews	320,023	9,765	6,280
MIND (Wu et al., 2020)	News	Sequence recommendation	15M	1M	160k

Table 3: A list of commonly used datasets in personalized LLMs for recommendation and search task. For the fifth column, the instances include reviews and ratings.

Big Five personality traits labels. Han et al. (2024) propose a Korean synthetic dataset with an emphasis on a Big Five personality-generating pipeline.

Takeaway. In this section, we survey various approaches for (1) generating personalized task-oriented responses and (2) modeling users personas. Currently, LLMs demonstrate a remarkable ability to model task-oriented dialogues and personas compared to many specialized models. However, there are potential exploits in persona modeling, as discussed in (§2.3). For example, Jin et al. (2024a); Shah et al. (2023); Deshpande et al. (2023) show that by assigning personas to LLMs, they can potentially aid in jailbreaking or produce toxic outputs. Further discussion on the challenges is outlined in (§5).

3.1.2 Recommendation System

In this section, we introduce the field of recommendation systems, a significant research area in personalization. Recommendation systems aim to recommend items to users that match their preferences. For example, systems may recommend various items such as books, movies, and restaurants. Moreover, we compile commonly used datasets in Table 3 and an overview of existing research in Table 4.

Existing studies explore various prompting methods for using LLMs in recommendation systems. Li et al. (2023c, 2021) develop a method to efficiently incorporate a user's personal information into pre-trained Transformer models. By using user and item ID information, they construct a recommendation system that outperforms others on the TripAdvisor, Amazon, and Yelp datasets. Li et al. (2023d) use LLMs for aspect extraction through prompt tuning. They combine aspect extraction with aspect-based recommendations. Chen et al. (2022a) propose a method to generate personalized chit-chat to enhance recommendation systems.

Paper	Scene	Dataset	Method	Task
Li et al. (2023d)	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., 2023c)	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. (2023a)	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. (2022a)	News	MIND (Wu et al., 2020), Reddit	Fine-tuning with weak labels	Direct Recommendation

Table 4: An overview of existing research in recommendation. Following the classification of Liu et al. (2023a), we classify recommendation systems into five types: rating prediction, sequential recommendation, explanation Generation, and review generation, and direct recommendation.

Wang and Lim (2023) utilize a three-step prompting process to achieve better performance.

Additionally, existing studies explore different frameworks for recommendation systems. Geng et al. (2022) unify various recommendation systems into a shared framework suitable for multiple downstream tasks. Yang et al. (2023b) fine-tune LLMs to perform recommendation tasks, producing a novel framework for building recommendation systems. 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.

Zero-Shot Recommendation System. In addition to the above research, some works focus on zero-shot recommendation systems Wang and Lim (2023); Hou et al. (2024). By utilizing in-context learning, Hou et al. (2024) propose a zero-shot sequential recommendation system. Furthermore, they show that LLMs in recommendation systems have biases due to popularity or item positions in the prompts. Zhang et al. (2023) design recommendation systems with enhanced user-friendliness, allowing users to interact freely with the system and receive more precise recommendations through natural language instructions. For generalizability, Wang et al. (2024f) highlights 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.

Takeaway. LLM-based recommendation systems use users' interaction histories to recommend specific items. Several works involve prompting methods, fine-tuning LLMs, or constructing new frameworks. Additionally, Chen (2023) provide a survey of LLM recommendation systems.

3.1.3 Personalized Search

In this section, we introduce applications of LLMs in personalized search systems. Traditional search systems provide a list of relevant search results based on input queries, often limited to simple queries. These results comprise various sources of information, which can be hard to distinguish and organize. Compared to traditional search systems, personalized search systems are based on complex input queries and past interactions to infer user preferences. They synthesize information from multiple sources and present the results 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. Therefore, personalized search has gained popularity as a research field.

There exists a line of investigation focusing on cognitive memory mechanism-based search systems. Zhou et al. (2021b) propose a method aimed at addressing the issue of data sparsity by using a contrastive sampling method to pre-train the ranking model. Furthermore, Zhou et al. (2021a) integrates LLMs with a cognitive memory mechanism, creating a model capable of accessing recent history without processing the entire user history. Building on this, Zhou et al. (2024) presents a strategy to combine a cognitive memory mechanism with LLMs to solve the problem of deep learning-based personalized search, which requires a large amount of training data. This enables LLMs to efficiently retrieve memory.

Empirically, Sharma et al. (2024) conducts experiments to investigate how LLM-powered search systems could lead to opinion polarization. Baek et al. (2024); Salemi and Zamani (2024) leverage search engine results to enhance LLM personalization.

Takeaway. LLM-based personalized search systems still require further development to ensure that LLMs are efficient and trustworthy with the search results. Li et al. (2024d) provide a comprehensive investigation of LLM-based search and recommendation systems.

3.1.4 Personalized Healthcare

LLMs for the Biomedical Domain. LLMs have exhibited expert-level capabilities in general biomedical domain tasks, including medical text generation (Hendrycks et al., 2021), medical question answering (Jin et al., 2020, 2019), and medical document classification (Cohan et al., 2020). Multi-modality models for the biomedical domain have the potential to integrate into people's everyday lives (Milne-Ives et al., 2020; Abbasian et al., 2024b). As healthcare assistance varies from person to person, a personalized health wellness assistant possesses a database of users' preferences and medical history, enabling it to provide precise medical advice or diagnoses.

Personalized LLMs for Healthcare. Abbasian et al. (2024a) propose OPENCHA, a framework for personalized healthcare that integrates knowledge from external data sources and personalized health data, and processes it using machine learning modules 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) combines parameter-efficient fine-tuning (Ding et al., 2023) (PEFT) with a memory retrieval module to generate personalized medical responses. Specifically, they separate the memory module into short-term memory (STM), which stores recent knowledge, and long-term memory (LTM), which retains frequently used knowledge from STM.

Other healthcare frameworks have been proposed. For example, HealthLLM (Jin et al., 2024b) combines feature extraction and knowledge retrieval, and then makes diagnosis predictions based on LlamaIndex (Liu, 2022) with XGBoost (Chen and Guestrin, 2016). HEALTHLLM (Jin et al., 2024b) is capable of generating personalized medical advice based on symptom descriptions provided by users.

Moreover, LLMs also show exceptional potential for psychotherapy (Stade et al., 2024; Chen et al., 2023b; Xu et al., 2024). In the near future, personalized LLMs for healthcare could potentially be integrated into clinical scenarios, mental health assessments, and prescribed therapeutic treatments.

Takeaway. As we witness rapid technological innovation in AI and its broader applications, several questions highlighted in Swift and Allen (2010) need to be considered to design a system that is both reliable and safe. Key issues include clinical decision-making involving AI and humans, and legal considerations concerning the liability associated with these personalized medical assistants.

3.1.5 Personalized Education

LLMs for Education. LLMs have demonstrated impressive abilities across a broad range of subjects from social sciences to natural sciences, encompassing topics like natural science (Wang et al., 2024d), coding (Haluptzok et al., 2023), mathematical reasoning (Ahn et al., 2024), medicine (Saab et al., 2024), finance (Yang et al., 2023c), and law (Cui et al., 2023). With the depth and breadth of knowledge compressed within LLMs, educators can harness their expertise to facilitate education in several ways. 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). Additionally, LLMs can assist in content creation to help educators develop more diverse and comprehensive materials (Xiao et al., 2023; Heck and Meurers, 2023), and implement automatic grading to ensure consistent grading standards (Matelsky et al., 2023; Tornqvist et al., 2023).

Personalized LLMs for Education. Recent surveys (Kasneci et al., 2023; Gan et al., 2023; Wang et al., 2024c; Jeon and Lee, 2023; Huber et al., 2024) 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. Furthermore, the integration of experts in various subjects to create a unified educator holds tremendous potential for the development of educational LLMs (Li et al., 2023e).

For example, EDUCHAT (Dan et al., 2023) pretrains models on an educational corpus, including textbooks, Chinese poetry, and psychology books, to establish a foundational knowledge base. Then, fine-tune the models on specialized tasks such as essay assessment, Socratic teaching data, emotional support, and retrieval-augmented open-question answering. HUMSUM (Shehata et al., 2023) summarizes 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.

Takeaway. Personalized LLMs have the potential to transform the educational ecosystem into one that is more inclusive and equitable, obviating the need for individuals to pay disproportionate fees. Individuals can access personalized educational content, assist teachers with lecture materials, and receive affordable tutoring, benefitting minority groups, underrepresented individuals, or those with limited access to resources. This concept of personalized LLMs is underscored by Tanya Byron's statement: "The technology itself is not transformative. It's the school, the pedagogy, that is transformative." In this context, it is LLMs for personalized education that will bring about the next transformative era.

3.2 LLM Personalized Methods

In this section, we introduce three major methods for personalizing LLMs. Additionally, beyond the categories discussed below, some works are difficult to classify. For example, to achieve customized preference learning, Cheng et al. (2023) propose a three-stage customized reward model learning process. Rocca and Yarkoni (2022) propose a user-encoding method by training with contrastive learning loss. Chu et al. (2023) train a recommendation model by applying a new mask mechanism, span order, positional encoding, and a dynamic position mechanism.

3.2.1 Fine-Tuning Method

Reinforcement Learning from Human Feedback (RLHF; Bai et al., 2022) is the dominant method for aligning LLMs with human intent. Several works have incorporated RLHF for personalizing LLMs. Mondal et al. (2024) propose a method of fine-tuning to create personalized slides from specific research papers. Hwang et al. (2023) use a multi-objective reinforcement learning (MORL) method to personalize robotic agents. Additionally, Jang et al. (2023) show that reframing the task of aligning LLMs with human preferences as a MORL problem can adapt to individual users better, outperforming supervised fine-tuning, RLHF, and prompting-based approaches. Li et al. (2024c) explain the deficit of vanilla RLHF and propose an advanced RLHF algorithm to fine-tune LLM. Kirk et al. (2023) also mention the drawback of RLHF that it may lead to a more polarised view.

In addition to RLHF, some works fine-tune LLMs using different methods: Shea and Yu (2023) propose an offline reinforcement learning method to enhance the personality consistency of dialogue systems. Qin et al. (2023) proposes to use data selection metrics and mini-batch data for more efficient LLM fine-tuning.

3.2.2 Retrieval Augmentation

For retrieval augmentation of personalizing LLMs, Wu et al. (2021) propose an attention-based split memory network. They use a splitting memory network of user profiles and users' comment histories on Reddit to generate personalized responses. Fu et al. (2022) trained a probabilistic model to select related personal memory as well as knowledge to generate personalized responses. Xu et al. (2022) propose a long-term memory mechanism to continuously extract user persona memory and agent persona memory from dialog history. Zhong et al. (2024) proposed MemoryBank, a long-term memory mechanism focusing on event summary and user portrait. Sun et al. (2024) propose a way to better represent data to make retrieval augmentation more efficient.

Limitations. Tan et al. (2024) consider that retrieval augmentation might underperform due to unrelated prompts. They propose a fine-tuning way to encode personalized parameters into LLMs. Prior studies have investigated memory-augmented techniques for prompting the LLM with pre-stored user-specific information to generate personalized responses for new queries. However, Zhang et al. (2024b) argues that this approach fails to capture fine-grained details.

3.2.3 Prompting Method

In this section, we introduce the prompting method to personalize LLMs. The prompting method is also the primary way to personalize LLMs. Following (Tan et al., 2024), we categorize prompt-based personalization methods fall into three categories.

Vanilla Personalized Prompt. For vanilla personalized Prompt, they directly prompt user preference and the required task to LLM. Dai et al. (2023) enhance the LLM's recommendation performance with traditional information retrieval (IR) ranking capabilities, including point-wise, pair-wise, and list-wise ranking. Christakopoulou et al. (2023) propose a user journeys method. LLM improves the recommendation system by reasoning through user activities. Additionally, Zhiyuli et al. (2023) employs a collaborative filtering algorithm to construct a book recommendation system.

Retrieval-augmented Personalized Prompt. For retrieval-augmented personalized prompts, prompt the LLM to effectively retrieve the relevant content. Mysore et al. (2023) propose a training method for prompt augmentation. Their training method makes LLM able to generate personalized social media posts. Salemi et al. (2024) Offer a comprehensive evaluation framework. Propose two retrieval augmentation approaches. (i) in-prompt augmentation (ii) fusion-in-decoder.i.e., encode personal items separately and integrate them in the decoder. Li et al. (2023a) personalizes LLMs by employing a method inspired by writing education. Specifically, they utilize the following steps to enhance LLMs' generation ability: multiple stages including retrieval, ranking, summarization, synthesis, and generation.

Profile-Augmented Prompt. Richardson et al. (2023) notes that incorporating user history data into the prompt to personalize LLMs can lead to inputs exceeding limitations and cost issues. Additionally, retrieval-based methods have the problem of potential information loss. Therefore, they propose a task-aware user summaries method to enable LLMs to achieve a deeper understanding of task profiles. Liu et al. (2023b) concentrates on content-based recommendation and examines both open-source and closed-source LLMs. They fine-tune the model for open-source LLMs and employ the prompting method for closed-source LLMs. Li et al. (2024a) train a prompt rewriter to improve prompt for better-personalized text generation.

4 LLM Personality Evaluation

In the previous sections, we summarize current progress in role-playing and personalized LLMs. Equivalently important is the research about quantifying and evaluating LLMs, like personality traits, and psychological behaviors. Understanding correct evaluation techniques allows us to better align LLM with the assigned role-playing model.

The most straightforward evaluation for LLM role-playing ability is personality. Some work evaluating LLMs with human personality evaluation, like Big Five (Jiang et al., 2023; Sorokovikova et al., 2024) and MBTI (Pan and Zeng, 2023; Song et al., 2024). There are arguments about the suitability to simply "transfer" human psychometric tests to LLMs (Dorner et al., 2023). Other works on creating a systematic and quantitative evaluation inventory based on Big Five tailored for LLMs (Jiang et al., 2024). Jiang et al. (2023); Sorokovikova et al. (2024) evaluate LLM personality based on the Big Five Personality Inventory (BFI) test[2] and story writing test.

In the BFI evaluation, LLMs often can reflect their assigned persona accurately. Moreover, their persona often influences their linguistic style and personality consistency (Frisch and Giulianelli, 2024; Jiang et al., 2023). While previous work solely focuses either on semantic accurate output and personality consistent dialogue, Harrison et al. (2019) explore simultaneously controlling semantic accuracy and personality consistency. Jiang et al. (2024) introduce Machine Personality Inventory (MPI) for evaluating LLMs' personality traits. They use Big Five Personality Factors to evaluate each personality trait consisting of a series of descriptions and a set of options and measure the mean and standard deviation of each trait. By comparing with human evaluation they found out that internal consistency (standard deviation) correlates with model capability.

Pan and Zeng (2023) evaluate LLM with the MBTI test measuring whether LLM possesses a human-like personality. Based on MBTI test results, they conclude that different LLMs possess different MBTI types, and are often affected by their training corpus. Moreover, using simple prompt-tuning have a hard time changing the MBTI type of LLMs. Wang et al. (2024e) evaluate role-playing agents *personality fidelity* by interviewing character agents for personality tests. Then based on the interview results they ask LLM to rate the score of each personality dimension. They found out that based on their measured personalities it matches well with the assigned character's personality traits. Song et al. (2024) propose an alternative evaluation method, instead of asking LLM questions and evaluating in Likert scale.

LLM Psychological Evaluation. In Huang et al. (2023), they recognize personality being only one aspect of psychological evaluation, they propose a framework PSYCHOBENCH for evaluating various psychological aspects of LLMs including: *personality traits* (i.e. Big Five Inventory), *interpersonal relationships*, *motivational tests*, *emotional abilities*. The results of LLMs personality agree with Jiang et al. (2024); Pan and Zeng (2023) that different LLMs exhibit distinct personality traits.

Potential Issues. Dorner et al. (2023) argue human personality evaluation (*e.g.* Big Five) cannot simply "transfer" to LLM. They experimentally show that *Agree Bias* (tendency to agree with both true and false key items, regardless of the actual item content) of LLM like GPT-4, GPT-3.5, and Llama-2 is significantly higher than that observed in human respondents. Other works (Song et al., 2024; Jiang et al., 2024) also show that LLMs sometimes do not have consistent personalities like humans. Additionally, Fang et al. (2023) addressed challenges in test-based personality computing research, including personality taxonomies, measurement quality, datasets, performance evaluation, and modeling choice. However, the question of whether it is possible to adopt human personality evaluation or do we need to design alternative test specifically for LLMs remains open.

LLMs as Evaluator. At the core of LLMs, the evaluator is replacing human or standardized metrics with LLMs as an evaluator. In Rao et al. (2023) they use LLM as a generic MBTI personality evaluator. They construct prompts by combining with original statements from MBTI questions with randomly permute options for unbiased reasons. In PSYCHOGAT (Yang et al., 2024), they use LLM-based gamified assessments to make psychological

[2] Big Five Personality Factors (John et al., 1999) are **O**penness, **C**onscientiousness, **E**xtroversion, **A**greeableness, and **N**euroticism.

diagnostics instead of based on standardized scales. The effectiveness of LLMs as personality evaluators shows impressive ability if provided with CoT prompting (Ji et al., 2023). On the other end, there are specialized LLMs as evaluators for evaluating other LLMS: PANDALM (Wang et al., 2023c) to distinguish the superior model from several LLMs; AUTO-J (Li et al., 2023b) for evaluating alignment of the models on *generality*, *flexibility*, and *generality*; CRITIQUELLM (Ke et al., 2023) aim at provide an explainable evaluation. They generate their training data from GPT-4 to obtain referenced and reference-free evaluation data, and fine-tuning CRITIQUELLM to generate the critiques.

Takeaway. To this day, there is still no unified and widely accepted way of precisely quantifying personality in LLMs. Do we evaluate LLMs the same as we evaluate human personality? If not, what are the alternative evaluation standards? What are some social and psychological impacts of different LLMs personalities? This will require further research as it is a relatively new area. As the capacity of LLMs increases beyond our imagination, it is crucial to develop a systematic measure to assess the psychological aspects of these LLMs. Just as personality and psychological traits are important predictors of a wide range of consequential life outcomes for humans, including physical and mental health, career success, social relationships, and major life trajectories (Roberts et al., 2007; Ozer and Benet-Martinez, 2006; Judge et al., 1999; McCrae and John, 1992; Hampson, 2012). It is crucial to measure personality and psychological traits in LLMs. This could be vital in the future as they take on more advanced roles and capabilities in society.

5 Challenges and Future Directions

5.1 Safety, Privacy, and Bias Issues

With the rapid development of LLM role-playing and LLM personalization, current studies have highlighted several vital yet often neglected issues that should be taken into consideration. In this section, we outline three issues: safety, privacy, and bias, advocating for the research community to further study these aspects in alignment with human values and to provide more reliable research (Si et al., 2022).

Safety. Several researchers have explored methods to bypass safety alignment, including *CipherChat* (Yuan et al., 2023) and Prompt Injection (Greshake et al., 2023). In particular, Chen et al. (2023c) and Li et al. (2024b) demonstrate the negative behaviors in LLM role-playing. Moreover, Jin et al. (2024a) and Shah et al. (2023) successfully manipulate LLMs to collaboratively jailbreak. Another safety issue is toxicity. Deshpande et al. (2023) presents that LLMs consistently exhibit toxicity about a range of topics when assigned personas. All of these works demonstrate the discovery of unsafe problems, indicating a need for more efforts to solve these.

Privacy. Since LLM personalization heavily relies on users' personas, including personal information and historical behaviors, privacy is a top priority in this field. Recently, Wang et al. (2024a) discovered that using the Membership Inference Attack can leak personal information in the training data, raising concerns about encoding personal data into models. Although several 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; Zanella-Béguelin et al., 2020), the research community still needs to put more effort into analyzing privacy issues and developing more acceptable personalized LLMs.

Bias. While a large number of studies focus on enhancing task performance by LLM role-playing, fewer works explore the disadvantages of assigning personas to 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. However, there is still ample room for investigating and mitigating bias issues in the field.

5.2 Lack of Datasets and Benchmarks

Several tasks in LLM role-playing require specific data formats or environmental information. For instance, web environments require HTML format, while game environments need settings and tools. Additionally, user persona modeling in personalized dialogue lacks contradictory persona datasets, which would more accurately represent real human behaviors (Kim et al., 2024b). Furthermore, LLM personalization faces a scarcity of high-quality personal data for model training due to privacy concerns. Although several personalization methods may exist, researchers are unable to thoroughly evaluate their performance. On the other hand, existing benchmarks for both LLM role-playing and LLM personalization are limited and lack comprehensive evaluations across various dimensions (Chang et al., 2023a). Therefore, expanding datasets and benchmarks for specialized usage and personal information under privacy protection is a potential future direction.

5.3 Multilingual and Multimodal Framework

Recent studies have proposed frameworks that harness the capabilities of LLMs for role-playing and personalization. We suggest extending these works into multilingual and multimodal settings. In terms of the multilingual aspect, although there are some multilingual persona datasets, little research has been conducted to experiment with and develop frameworks using them. As knowledge transformation between languages is vital (Tanwar et al., 2023), multilingual frameworks could be a potential future direction. Regarding the multimodal aspect, several studies have been conducted in web environments leveraging LLMs, where models can utilize both text and visual information (Koh et al., 2024; Liu et al., 2024). However, for other environments and LLM personalization, there is still a lack of these kinds of multimodal frameworks for better utilizing LLM capabilities.

5.4 Connecting to Psychology Research

LLMs have shown tremendous potential, however, whether LLMs pose an existential risk to humankind still requires further investigations. An increasingly popular trend of such evaluation is to draw inspiration from the Psychology field. One of the most comprehensive evaluation methods is PsychoBench (Huang et al., 2023), which comprises four aspects: personality traits, interpersonal relationships, motivational tests, and emotional abilities. Studies by Jiang et al. (2023) and Sorokovikova et al. (2024) present that LLMs exhibit distinct personality traits. Nonetheless, as discussed in (§4), there is currently no unified and widely accepted method to quantify the evaluation of LLMs. Thus, conducting comprehensive psychological research on the personality of LLMs, or even on persona-assigned LLMs, could be a meaningful direction for future research.

6 Conclusion

Leveraging personas, LLMs can generate tailored responses, effectively adapting to a wide range of scenarios. In this survey paper, we outline two lines of work – role-playing and personalization – for research of personas in the era of LLMs. We first present LLM role-playing, introducing their agentic schemas and emergent behaviors under various interactive environments. We then elaborate on LLM personalization, categorizing different tasks and associated methods for user alignment. Lastly, we highlight current challenges and promising future directions.

To the best of our knowledge, we present the first survey, under the unified view of persona, for LLMs role-playing and personalization. We hope this work serves as an introductory guide for beginners to the field and a stepping stone to foster future endeavors for the community.

Acknowledgements

This work was financially supported by the National Science and Technology Council (NSTC) in Taiwan, under Grants 112-2223-E-002-012-MY5 and 111-2222-E-002-013-MY3, and from the Featured Area Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (113L900901/113L900902/113L900903).

References

Mahyar Abbasian, Iman Azimi, Amir M. Rahmani, and Ramesh Jain. 2024a. Conversational health agents: A personalized llm-powered agent framework. 3, 12

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. 12

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. 3, 12

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*. 5

Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. 2024. Large language models for mathematical reasoning: Progresses and challenges. 13

Md Adnan Arefeen, Biplob Debnath, and Srimat Chakradhar. 2023. Leancontext: Cost-efficient domain-specific question answering using llms. 13

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. 9

Jinheon Baek, Nirupama Chandrasekaran, Silviu Cucerzan, Allen herring, and Sujay Kumar Jauhar. 2024. Knowledge-augmented large language models for personalized contextual query suggestion. 12

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. 14

Namo Bang, Jeehyun Lee, and Myoung-Wan Koo. 2023. Task-optimized adapters for an end-to-end task-oriented dialogue system. *arXiv preprint arXiv:2305.02468*. 9

Youssef Bassil. 2012. A simulation model for the waterfall software development life cycle. *arXiv preprint arXiv:1205.6904*. 4

R Meredith Belbin and Victoria Brown. 2022. *Team roles at work*. Routledge. 4

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. 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. 3, 9, 10

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*. 3, 6, 7

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. 2023a. A survey on evaluation of large language models. 17

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023b. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*. 6

Changyu Chen, Xiting Wang, Xiaoyuan Yi, Fangzhao Wu, Xing Xie, and Rui Yan. 2022a. 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. 10, 11

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

Junyi Chen. 2023. A survey on large language models for personalized and explainable recommendations. 12

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*. 4

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*. 13

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16. ACM. 13

Wei Chen, Zhiwei Li, Hongyi Fang, Qianyuan Yao, Cheng Zhong, Jianye Hao, Qi Zhang, Xuanjing Huang, Jiajie Peng, and Zhongyu Wei. 2022b. A benchmark for automatic medical consultation system: Frameworks, tasks and datasets. 3

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*. 3, 6, 7, 16

Yang Chen, Ethan Mendes, Sauvik Das, Wei Xu, and Alan Ritter. 2023d. Can language models be instructed to protect personal information? 16

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. 7

Kanzhi Cheng, Qiushi Sun, Yougang Chu, Fangzhi Xu, Yantao Li, Jianbing Zhang, and Zhiyong Wu. 2024a. Seeclick: Harnessing gui grounding for advanced visual gui agents. *arXiv preprint arXiv:2401.10935*. 3, 5

Pengyu Cheng, Jiawen Xie, Ke Bai, Yong Dai, and Nan Du. 2023. Everyone deserves a reward: Learning customized human preferences. 13

Yuheng Cheng, Ceyao Zhang, Zhengwen Zhang, Xiangrui Meng, Sirui Hong, Wenhao Li, Zihao Wang, Zekai Wang, Feng Yin, Junhua Zhao, et al. 2024b. Exploring large language model based intelligent agents: Definitions, methods, and prospects. *arXiv preprint arXiv:2401.03428*. 7

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*. 6

Konstantina Christakopoulou, Alberto Lalama, Cj Adams, Iris Qu, Yifat Amir, Samer Chucri, Pierce Vollucci, Fabio Soldo, Dina Bseiso, Sarah Scodel, Lucas Dixon, Ed H. Chi, and Minmin Chen. 2023. Large language models for user interest journeys. 3, 14

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 pretrained recommender systems. 11, 13

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. 12

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases. 13

Sunhao Dai, Ninglu Shao, Haiyuan Zhao, Weijie Yu, Zihua Si, Chen Xu, Zhongxiang Sun, Xiao Zhang, and Jun Xu. 2023. Uncovering chatgpt's capabilities in recommender systems. In *Proceedings of the 17th ACM Conference on Recommender Systems*, RecSys '23. ACM. 3, 14

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*. 3, 13

Tom DeMarco and Tim Lister. 2013. *Peopleware: productive projects and teams*. Addison-Wesley. 4

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. 3, 5

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. 7

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. 7, 10, 16

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). 3, 7, 10

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. 2023. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3):220–235. 12

Yihong Dong, Xue Jiang, Zhi Jin, and Ge Li. 2023. Self-collaboration code generation via chatgpt. *arXiv preprint arXiv:2304.07590*. 3, 4

Florian E Dorner, Tom Sühr, Samira Samadi, and Augustin Kelava. 2023. Do personality tests generalize to large language models? *arXiv preprint arXiv:2311.05297*. 15

Yiming Du, Hongru Wang, Zhengyi Zhao, Bin Liang, Baojun Wang, Wanjun Zhong, Zezhong Wang, and Kam-Fai Wong. 2024. Perltqa: A personal long-term memory dataset for memory classification, retrieval, and synthesis in question answering. 3

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. 9, 10

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. 15

Yihao Feng, Shentao Yang, Shujian Zhang, Jianguo Zhang, Caiming Xiong, Mingyuan Zhou, and Huan Wang. 2023. Fantastic rewards and how to tame them: A case study on reward learning for task-oriented dialogue systems. *arXiv preprint arXiv:2302.10342*. 9

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. 15

Tingchen Fu, Xueliang Zhao, Chongyang Tao, Ji-Rong Wen, and Rui Yan. 2022. There are a thousand hamlets in a thousand people's eyes: Enhancing knowledge-grounded dialogue with personal memory. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3901–3913, Dublin, Ireland. Association for Computational Linguistics. 3, 14

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, 5, 7

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. 16

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. 13

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. 11

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. 13

Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. 2023. Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. 16

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, 7

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*. 7, 16

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*. 5

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*. 5

Patrick Haluptzok, Matthew Bowers, and Adam Tauman Kalai. 2023. Language models can teach themselves to program better. 13

Sarah E Hampson. 2012. Personality processes: Mechanisms by which personality traits "get outside the skin". *Annual review of psychology*, 63:315–339. 16

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. 10

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. 13

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. 3, 10, 11

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. 15

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. 10

Tanja Heck and Detmar Meurers. 2023. Using learning analytics for adaptive exercise generation. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 44–56, Toronto, Canada. Association for Computational Linguistics. 13

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. 12

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*. 3, 4, 7

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*. 3, 5

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. 8, 9

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. 11

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. 11

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. 3, 9

Jen-tse Huang, Wenxuan Wang, Eric John Li, Man Ho LAM, Shujie Ren, Youliang Yuan, Wenxiang Jiao, Zhaopeng Tu, and Michael Lyu. 2023. On the humanity of conversational ai: Evaluating the psychological portrayal of llms. In *The Twelfth International Conference on Learning Representations*. 15, 17

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. 16

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. 8

Xu Huang, Weiwen Liu, Xiaolong Chen, Xingmei Wang, Hao Wang, Defu Lian, Yasheng Wang, Ruiming Tang, and Enhong Chen. 2024. Understanding the planning of llm agents: A survey. *arXiv preprint arXiv:2402.02716*. 2

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. 13

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. 9

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. 9

Minyoung Hwang, Luca Weihs, Chanwoo Park, Kimin Lee, Aniruddha Kembhavi, and Kiana Ehsani. 2023. Promptable behaviors: Personalizing multi-objective rewards from human preferences. 14

Léo Jacqmin, Lina M. Rojas-Barahona, and Benoit Favre. 2022. "do you follow me?": A survey of recent approaches in dialogue state tracking. 8

Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. 2023. Personalized soups: Personalized large language model alignment via post-hoc parameter merging. *arXiv preprint arXiv:2310.11564*. 14

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. 13

Yu Ji, Wen Wu, Hong Zheng, Yi Hu, Xi Chen, and Liang He. 2023. Is chatgpt a good personality recognizer? a preliminary study. 16

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. 15

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*. 15, 17

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. 12

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*. 10, 16

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. 3, 12, 13

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. 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 2567–2577, Hong Kong, China. Association for Computational Linguistics. 12

Oliver P John, Sanjay Srivastava, et al. 1999. The big-five trait taxonomy: History, measurement, and theoretical perspectives. 15

Timothy A Judge, Chad A Higgins, Carl J Thoresen, and Murray R Barrick. 1999. The big five personality traits, general mental ability, and career success across the life span. *Personnel psychology*, 52(3):621–652. 16

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. 13

Pei Ke, Bosi Wen, Zhuoer Feng, Xiao Liu, Xuanyu Lei, Jiale Cheng, Shengyuan Wang, Aohan Zeng, Yuxiao Dong, Hongning Wang, Jie Tang, and Minlie Huang. 2023. Critiquellm: Scaling llm-as-critic for effective and explainable evaluation of large language model generation. 16

Geunwoo Kim, Pierre Baldi, and Stephen McAleer. 2024a. Language models can solve computer tasks. *Advances in Neural Information Processing Systems*, 36. 4

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*. 3, 16

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*. 8

Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A. Hale. 2023. Personalisation within bounds: A risk taxonomy and policy framework for the alignment of large language models with personalised feedback. 14

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*. 3, 4, 5, 17

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, and Xin Zhou. 2023. Better zero-shot reasoning with role-play prompting. *arXiv preprint arXiv:2308.07702*. 2

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. 8

Mike Lewis, Denis Yarats, Yann N Dauphin, Devi Parikh, and Dhruv Batra. 2017. Deal or no deal? end-to-end learning for negotiation dialogues. *arXiv preprint arXiv:1706.05125*. 5

Cheng Li, Mingyang Zhang, Qiaozhu Mei, Weize Kong, and Michael Bendersky. 2024a. Learning to rewrite prompts for personalized text generation. In *Proceedings of the ACM on Web Conference 2024*, WWW '24. ACM. 15

Cheng Li, Mingyang Zhang, Qiaozhu Mei, Yaqing Wang, Spurthi Amba Hombaiah, Yi Liang, and Michael Bendersky. 2023a. Teach llms to personalize – an approach inspired by writing education. 3, 14

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2024b. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36. 2, 3, 6, 7, 16

Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu. 2023b. Generative judge for evaluating alignment. *arXiv preprint arXiv:2310.05470*. 16

Lei Li, Yongfeng Zhang, and Li Chen. 2021. Personalized transformer for explainable recommendation. *arXiv preprint arXiv:2105.11601*. 10, 11

Lei Li, Yongfeng Zhang, and Li Chen. 2023c. Personalized prompt learning for explainable recommendation. *ACM Transactions on Information Systems*, 41(4):1–26. 10, 11

Pan Li, Yuyan Wang, Ed H. Chi, and Minmin Chen. 2023d. Prompt tuning large language models on personalized aspect extraction for recommendations. 10, 11

Qingyao Li, Lingyue Fu, Weiming Zhang, Xianyu Chen, Jingwei Yu, Wei Xia, Weinan Zhang, Ruiming Tang, and Yong Yu. 2023e. Adapting large language models for education: Foundational capabilities, potentials, and challenges. *arXiv preprint arXiv:2401.08664*. 13

Xinyu Li, Zachary C. Lipton, and Liu Leqi. 2024c. Personalized language modeling from personalized human feedback. 3, 14

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. 10

Yongqi Li, Xinyu Lin, Wenjie Wang, Fuli Feng, Liang Pang, Wenjie Li, Liqiang Nie, Xiangnan He, and Tat-Seng Chua. 2024d. A survey of generative search and recommendation in the era of large language models. 12

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023f. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. 3

Zekun Li, Baolin Peng, Pengcheng He, Michel Galley, Jianfeng Gao, and Xifeng Yan. 2024e. Guiding large language models via directional stimulus prompting. *Advances in Neural Information Processing Systems*, 36. 3, 9

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

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81. 6

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. 6

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. 8

Jerry Liu. 2022. LlamaIndex. 13

Junling Liu, Chao Liu, Peilin Zhou, Renjie Lv, Kang Zhou, and Yan Zhang. 2023a. Is chatgpt a good recommender? a preliminary study. 11

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*. 3, 5, 17

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. 3, 9

Qijiong Liu, Nuo Chen, Tetsuya Sakai, and Xiao-Ming Wu. 2023b. Once: Boosting content-based recommendation with both open- and closed-source large language models. 3, 14

Ehsan Lotfi, Maxime De Bruyn, Jeska Buhmann, and Walter Daelemans. 2024. Personalitychat: Conversation distillation for personalized dialog modeling with facts and traits. 9

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2017. Hierarchical question-image co-attention for visual question answering. 9

Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. 2023. Analyzing leakage of personally identifiable information in language models. 16

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

Andrea Madotto, Zihan Liu, Zhaojiang Lin, and Pascale Fung. 2020. Language models as few-shot learner for task-oriented dialogue systems. *arXiv preprint arXiv:2008.06239*. 9

Jordan K Matelsky, Felipe Parodi, Tony Liu, Richard D Lange, and Konrad P Kording. 2023. A large language model-assisted education tool to provide feedback on open-ended responses. *arXiv preprint arXiv:2308.02439*. 13

Robert R McCrae and Oliver P John. 1992. An introduction to the five-factor model and its applications. *Journal of personality*, 60(2):175–215. 16

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. 9

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. 12

Ishani Mondal, Shwetha S, Anandhavelu Natarajan, Aparna Garimella, Sambaran Bandyopadhyay, and Jordan Boyd-Graber. 2024. Presentations by the humans and for the humans: Harnessing LLMs for generating persona-aware slides from documents. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2664–2684, St. Julian's, Malta. Association for Computational Linguistics. 14

Johannes E. M. Mosig, Shikib Mehri, and Thomas Kober. 2020. Star: A schema-guided dialog dataset for transfer learning. 10

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. 8

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*. 8

Nikola Mrkšić and Ivan Vulić. 2018. Fully statistical neural belief tracking. 8

Sheshera Mysore, Zhuoran Lu, Mengting Wan, Longqi Yang, Steve Menezes, Tina Baghaee, Emmanuel Barajas Gonzalez, Jennifer Neville, and Tara Safavi. 2023. Pearl: Personalizing large language model writing assistants with generation-calibrated retrievers. 3, 14

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. 10, 11

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

Daniel J Ozer and Veronica Benet-Martinez. 2006. Personality and the prediction of consequential outcomes. *Annu. Rev. Psychol.*, 57:401–421. 16

Keyu Pan and Yawen Zeng. 2023. Do llms possess a personality? making the mbti test an amazing evaluation for large language models. 15

PapersWithCode. 2020. Baidu personachat dataset. `https://paperswithcode.com/dataset/baidu-personachat`. 10

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318. 6

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, 5

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*. 3, 13

Greg Pass, Abdur Chowdhury, and Cayley Torgeson. 2006. A picture of search. In *Proceedings of the 1st international conference on Scalable information systems*, pages 1–es. 3

Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao. 2021. Soloist: Building task bots at scale with transfer learning and machine teaching. *Transactions of the Association for Computational Linguistics*, 9:807–824. 9

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. 4

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, 7

Ruiyang Qin, Jun Xia, Zhenge Jia, Meng Jiang, Ahmed Abbasi, Peipei Zhou, Jingtong Hu, and Yiyu Shi. 2023. Enabling on-device large language model personalization with self-supervised data selection and synthesis. *arXiv preprint arXiv:2311.12275*. 14

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. 9, 10

Haocong Rao, Cyril Leung, and Chunyan Miao. 2023. Can chatgpt assess human personalities? a general evaluation framework. *arXiv preprint arXiv:2303.01248*. 15

Abhinav Rastogi, Dilek Hakkani-Tur, and Larry Heck. 2018. Scalable multi-domain dialogue state tracking. 8

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. 3, 10

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. 3, 14

Brent W Roberts, Nathan R Kuncel, Rebecca Shiner, Avshalom Caspi, and Lewis R Goldberg. 2007. The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological science*, 2(4):313–345. 16

Roberta Rocca and Tal Yarkoni. 2022. Language as a fingerprint: Self-supervised learning of user encodings using transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1701–1714, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. 13

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*. 13

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024. Lamp: When large language models meet personalization. 3, 14

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*. 12

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. 10, 16

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*. 12

Ryan Shea and Zhou Yu. 2023. Building persona consistent dialogue agents with offline reinforcement learning. 14

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. 3, 13

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

Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Boyd-Graber, and Lijuan Wang. 2022. Prompting gpt-3 to be reliable. *arXiv preprint arXiv:2210.09150*. 16

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. 2016. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489. 5

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. 2017. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359. 5

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. 5

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. 9

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. 15

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. 8

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. 15, 17

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. 12

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. 13

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. 10

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. 14

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. 8

M Swift and J Allen. 2010. Towards a personal health management assistant. *Journal of biomedical informatics*, 43(5):S13–S16. 13

Zhaoxuan Tan, Qingkai Zeng, Yijun Tian, Zheyuan Liu, Bing Yin, and Meng Jiang. 2024. Democratizing large language models via personalized parameter-efficient fine-tuning. 14

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, 5, 7

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. 3, 9

Eshaan Tanwar, Subhabrata Dutta, Manish Borthakur, and Tanmoy Chakraborty. 2023. Multilingual LLMs are better cross-lingual in-context learners with alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6292–6307, Toronto, Canada. Association for Computational Linguistics. 17

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*. 5

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940. 5

Maximilian Tornqvist, Mosleh Mahamud, Erick Mendez Guzman, and Alexandra Farazouli. 2023. ExASAG: Explainable framework for automatic short answer grading. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 361–371, Toronto, Canada. Association for Computational Linguistics. 13

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*. 9

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, 5, 6

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. 16

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*. 8

Lei Wang and Ee-Peng Lim. 2023. Zero-shot next-item recommendation using large pretrained language models. 11

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2024b. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):1–26. 2, 6

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023b. Large language models are not fair evaluators. 6

Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S Yu, and Qingsong Wen. 2024c. Large language models for education: A survey and outlook. *arXiv preprint arXiv:2403.18105*. 13

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R. Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. 2024d. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. 13

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. 2024e. Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. 15

Yancheng Wang, Ziyan Jiang, Zheng Chen, Fan Yang, Yingxue Zhou, Eunah Cho, Xing Fan, Xiaojiang Huang, Yanbin Lu, and Yingzhen Yang. 2024f. Recmind: Large language model powered agent for recommendation. 12

Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, et al. 2023c. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization. *arXiv preprint arXiv:2306.05087*. 16

Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023d. Aligning large language models with human: A survey. *arXiv preprint arXiv:2307.12966*. 6

Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2023e. Unleashing cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. *arXiv preprint arXiv:2307.05300*, 1(2):3. 2, 3, 6, 7

Zhilin Wang, Yu Ying Chiu, and Yu Cheung Chiu. 2023f. Humanoid agents: Platform for simulating human-like generative agents. *arXiv preprint arXiv:2310.05418*. 3, 5

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. 10

Bowen Wu, MengYuan Li, Zongsheng Wang, Yifu Chen, Derek Wong, Qihang Feng, Junhong Huang, and Baoxun Wang. 2019a. Guiding variational response generator to exploit persona. *arXiv preprint arXiv:1911.02390*. 9

Cheng-Kuang Wu, Wei-Lin Chen, and Hsin-Hsi Chen. 2023a. Large language models perform diagnostic reasoning. *arXiv preprint arXiv:2307.08922*. 2, 3, 5

Chien-Sheng Wu, Andrea Madotto, Ehsan Hosseini-Asl, Caiming Xiong, Richard Socher, and Pascale Fung. 2019b. 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. 8

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. 10, 11

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. 9

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. 3, 6

Qingyang Wu, James Gung, Raphael Shu, and Yi Zhang. 2023c. Diacttod: Learning generalizable latent dialogue acts for controllable task-oriented dialogue systems. *arXiv preprint arXiv:2308.00878*. 9

Yuwei Wu, Xuezhe Ma, and Diyi Yang. 2021. Personalized response generation via generative split memory network. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1956–1970, Online. Association for Computational Linguistics. 3, 14

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*. 2, 6, 7

Changrong Xiao, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Lei Xia. 2023. Evaluating reading comprehension exercises generated by LLMs: A showcase of ChatGPT in education applications. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 610–625, Toronto, Canada. Association for Computational Linguistics. 13

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. 9

Xinchao Xu, Zhibin Gou, Wenquan Wu, Zheng-Yu Niu, Hua Wu, Haifeng Wang, and Shihang Wang. 2022. Long time no see! open-domain conversation with long-term persona memory. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2639–2650, Dublin, Ireland. Association for Computational Linguistics. 3, 14

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. 13

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. 3, 9

Fan Yang, Zheng Chen, Ziyan Jiang, Eunah Cho, Xiaojiang Huang, and Yanbin Lu. 2023b. Palr: Personalization aware llms for recommendation. 3, 11

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023c. Fingpt: Open-source financial large language models. 13

Min Yang, Weiyi Huang, Wenting Tu, Qiang Qu, Ying Shen, and Kai Lei. 2021a. Multitask learning and reinforcement learning for personalized dialog generation: An empirical study. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):49–62. 8

Qisen Yang, Zekun Wang, Honghui Chen, Shenzhi Wang, Yifan Pu, Xin Gao, Wenhao Huang, Shiji Song, and Gao Huang. 2024. Llm agents for psychology: A study on gamified assessments. 15

Yunyi Yang, Yunhao Li, and Xiaojun Quan. 2021b. Ubar: Towards fully end-to-end task-oriented dialog system with gpt-2. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14230–14238. 9

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. 2022. Webshop: Towards scalable real-world web interaction with grounded language agents. *Advances in Neural Information Processing Systems*, 35:20744–20757. 5

Yelp. 2013. Yelp dataset. `https://www.yelp.com/dataset`. 10

Youliang Yuan, Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Pinjia He, Shuming Shi, and Zhaopeng Tu. 2023. Gpt-4 is too smart to be safe: Stealthy chat with llms via cipher. *arXiv preprint arXiv:2308.06463*. 16

Idris Yusupov and Yurii Kuratov. 2018. NIPS conversational intelligence challenge 2017 winner system: Skill-based conversational agent with supervised dialog manager. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3681–3692, Santa Fe, New Mexico, USA. Association for Computational Linguistics. 7

Santiago Zanella-Béguelin, Lukas Wutschitz, Shruti Tople, Victor Rühle, Andrew Paverd, Olga Ohrimenko, Boris Köpf, and Marc Brockschmidt. 2020. Analyzing information leakage of updates to natural language models. In *Proceedings of the 2020 ACM SIGSAC conference on computer and communications security*, pages 363–375. 16

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. 9, 10, 31

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, Hongchao Fang, Penghui Zhu, Shu Chen, and Pengtao Xie. 2020. MedDialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250, Online. Association for Computational Linguistics. 3

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. 8

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. 11

Kai Zhang, Yangyang Kang, Fubang Zhao, and Xiaozhong Liu. 2024a. Llm-based medical assistant personalization with short- and long-term memory coordination. 3, 12

Kai Zhang, Lizhi Qing, Yangyang Kang, and Xiaozhong Liu. 2024b. Personalized llm response generation with parameterized memory injection. 14

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? 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. 3, 9, 32

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? 9, 10

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*. 6

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*. 4, 5

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. 6

Yinhe Zheng, Rongsheng Zhang, Xiaoxi Mao, and Minlie Huang. 2019. A pre-training based personalized dialogue generation model with persona-sparse data. 9

Aakas Zhiyuli, Yanfang Chen, Xuan Zhang, and Xun Liang. 2023. Bookgpt: A general framework for book recommendation empowered by large language model. 3, 14

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. 9

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. 3, 14

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*. 3, 5

Yujia Zhou, Zhicheng Dou, Bingzheng Wei, and Ruobing Xievand Ji-Rong Wen. 2021a. Group based personalized search by integrating search behaviour and friend network. 12

Yujia Zhou, Zhicheng Dou, Yutao Zhu, and Ji-Rong Wen. 2021b. Pssl: Self-supervised learning for personalized search with contrastive sampling. 12

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*. 3, 12

Noah Ziems, Wenhao Yu, Zhihan Zhang, and Meng Jiang. 2023. Large language models are built-in autoregressive search engines. 12

A Personalized Dialogue Generation Example

A.1 ToD Modeling Example

In Table 5 we provide an example for Task-oriented Dialogue Modeling (§3.1.1).

A.2 User Persona Modeling Example

In Table 6 we provide and example for Persona Modeling Example (§3.1.1).

Table 5: **Dialogue Example from MultiWOZ Dataset (Zang et al., 2020).**

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 6: **User Persona Modeling Example from PERSONA-CHAT (Zhang et al., 2018a).**

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.