Personalization of Large Language Models: A Survey

Anonymous authors
Paper under double-blind review

Abstract

Personalization of Large Language Models (LLMs) has recently become increasingly important with a wide range of applications. Despite the importance and recent progress, most existing works on personalized LLMs have focused either entirely on (a) personalized text generation or (b) leveraging LLMs for personalization-related downstream applications, such as recommendation systems. In this work, we bridge the gap between these two separate main directions for the first time by introducing a taxonomy for personalized LLM usage and summarizing the key differences and challenges. We provide a formalization of the foundations of personalized LLMs that consolidates and expands notions of personalization of LLMs, defining and discussing novel facets of personalization, usage, and desiderata of personalized LLMs. We then unify the literature across these diverse fields and usage scenarios by proposing systematic taxonomies for the granularity of personalization, personalization techniques, datasets, evaluation methods, and applications of personalized LLMs. Finally, we highlight challenges and important open problems that remain to be addressed. By unifying and surveying recent research using the proposed taxonomies, we aim to provide a clear guide to the existing literature and different facets of personalization in LLMs, empowering both researchers and practitioners.

1 Introduction

Large language models (LLMs) have emerged as powerful tools capable of performing a wide range of natural language processing (NLP) tasks with remarkable proficiency (*e.g.*, Radford et al., 2018; Devlin et al., 2019; Lewis et al., 2019; Radford et al., 2019; Brown et al., 2020; Raffel et al., 2020; Achiam et al., 2023; Touvron et al., 2023; Groeneveld et al., 2024). Empirically, these models have demonstrated their capability as generalist models, allowing them to perform numerous tasks such as text generation, translation, summarization, and question-answering with decent accuracy. Notably, LLMs can perform effectively in zero-shot or few-shot settings, meaning they can follow human instructions and perform complex tasks with little to no task-specific training data (Bommasani et al., 2021; Liu et al., 2023c). This capability eliminates the need for extensive fine-tuning of their parameters, thereby significantly simplifying human interaction with machines through straightforward input prompts. For instance, users can engage with LLMs in a conversational format, making interactions more intuitive and accessible. Such robust and versatile abilities of LLMs have led to the creation of numerous applications, including general AI assistants (AutoGPT, 2024), copilots (Microsoft, 2024), and personal LLM-based agents (Li et al., 2024h). These applications assist users in a wide range of activities such as writing emails, generating code, drafting reports, and more.

Recently, there has been growing interest in adapting LLMs to user-specific contexts, beyond their natural use as NLP task solvers or general-purpose chatbots (Tseng et al., 2024). To this end, personalization of LLMs addresses this by adapting the models to generate responses that cater to the unique needs and preferences of each user or user group (Salemi et al., 2023). Such personalization is crucial for human-AI interaction and user-focused applications. It is expected to enhance user satisfaction by providing more relevant and meaningful interactions, ensuring users receive responses that are more aligned with their needs and expectations. This enables LLMs to offer more effective assistance across a diverse range of applications such as customer support (Amazon, 2024), where personalized responses can significantly improve user experience; education (Wang et al., 2022; 2024b), where tailored content can better meet individual learning

needs (Woźniak et al., 2024); and healthcare, where personalized advice can enhance patient care (Tang et al., 2023; Yuan et al., 2023).

Personalization of LLMs has recently attracted significant attention (Salemi et al., 2023; Tseng et al., 2024), with research primarily focusing on two directions: (a) personalized text generation, which tailors generated text to user-specific contexts, and (b) downstream task personalization, which leverages LLM capabilities to improve performance on targeted applications such as recommendation systems. Despite the extensive research efforts, these two areas have historically developed independently due to technical limitations and methodological differences, often resulting in existing surveys (Chen, 2023; Chen et al., 2024b;c) examining each aspect in isolation. However, these two domains are not fundamentally distinct, as LLMs possess the flexibility to adapt to a wide range of tasks (Qin et al., 2023). Rather, as LLM capabilities continue to evolve, we envision that those two directions will increasingly converge, enabling unified systems in which a single intelligent agent seamlessly transitions from engaging in personalized conversations to reasoning over structured knowledge like product catalogs for task-oriented recommendations. Bridging this current conceptual gap by synthesizing insights across both dimensions thus constitutes an essential step toward creating fully integrated, adaptable, and generalizable user experiences.

To fully understand LLM personalization, it is important to examine these research directions within a unified framework that captures the broader landscape of personalization. Beyond integrating different approaches, a more in-depth discussion is needed on the foundational concepts, techniques, datasets, and evaluation methods that support LLM personalization. Additionally, real-world challenges such as balancing personalization with privacy concerns, mitigating biases, and handling data limitations must be addressed to ensure practical and ethical deployment. In this survey, we provide a comprehensive perspective by introducing systematic taxonomies that categorize personalization based on granularity, methodology, and evaluation strategies. We also explore open problems and potential research opportunities. Through this framework, we connect the various aspects of personalization and offer a structured reference that outlines the essential components needed to develop and evaluate personalized LLMs.

The key contributions of this work are as follows:

1. **A unifying view and taxonomy for the usage of personalized LLMs (Section 2).** We provide a unifying view and taxonomy of the usage of personalized LLMs based on whether they focus on evaluating the generated text directly, or whether the text is used indirectly for another downstream application. This serves as a fundamental basis for understanding and unifying the two separate areas focused on the personalization of LLMs. Further, we analyze the limitations of each, including the features, evaluation, and datasets, among other factors.
2. **A formalization of personalized LLMs (Section 3).** We provide a formalization of personalized LLMs by establishing foundational concepts that consolidate existing notions of personalization, defining and discussing novel facets of personalization, and outlining desiderata for their application across diverse usage scenarios.
3. **An analysis and taxonomy of the personalization granularity of LLMs (Section 4).** We propose three different levels of personalization granularity for LLMs, including (i) user-level personalization, (ii) persona-level personalization, and (iii) global preference personalization. We formalize these levels, and then discuss and characterize the trade-offs between the different granularities of LLM personalization. Notably, user-level personalization is the finest granularity; however, it requires a sufficient amount of user-level data. In contrast, persona-level personalization groups users into personas and tailors the experience based on persona assignments. While it doesn't provide the same granularity as user-level personalization, it is effective for personalizing experiences for users with limited data. Finally, global personalization caters to the overall preferences of the general public and does not offer user-specific personalization.[1]
4. **A survey and taxonomy of techniques for LLM personalization (Section 5).** We categorize and provide a comprehensive overview of the current techniques for personalizing LLMs based on how user information is utilized. Our taxonomy covers various categories of methods such as retrieval-augmented generation (RAG), prompt engineering, supervised fine-tuning, embedding learning, and

[1]We include it here for completeness, though it is not the focus of this work.

reinforcement learning from human feedback (RLHF). For each category of methods, we discuss their unique characteristics, applications, and the trade-offs involved. Our detailed analysis helps in understanding the strengths and limitations of different personalization techniques and their suitability for various tasks.

5. **A survey and taxonomy of metrics and evaluation of personalized LLMs (Section 6).** We categorize and analyze the existing metrics used for evaluating personalized LLMs, proposing a novel taxonomy that distinguishes between direct and indirect evaluation methods. We highlight the importance of both qualitative and quantitative metrics, addressing various facets such as user satisfaction, relevance, and coherence of the generated text. Additionally, we discuss the challenges in evaluating personalized LLMs and suggest potential solutions to improve the robustness and reliability of the evaluation process.
6. **A survey and taxonomy of datasets for personalized LLMs (Section 7).** We provide a comprehensive taxonomy of datasets used for training and evaluating personalized LLMs, categorizing them based on their usage in direct or indirect evaluation of personalized text generation. Our survey covers a wide range of datasets, including those specifically designed for short- and long-text generation, recommendation systems, classification tasks, and dialogue generation. We discuss the strengths and limitations of each dataset, their relevance to different personalization techniques, and the need for more diverse and representative datasets to advance the field.
7. **A survey of applications for personalized LLMs (Section 8).** We survey key domains where personalized LLMs are applied, including AI assistants in education and healthcare, finance, legal, and coding environments. We also explore their use in recommendation systems and search engines, highlighting the ability of personalized LLMs to deliver customized user experiences, enhance engagement, and improve task-specific outcomes across diverse fields.
8. **An overview of important open problems and challenges for future work to address (Section 9).** We outline critical challenges and open research questions in personalized LLMs that need to be addressed for advancing the field. Key issues include the need for improved benchmarks and metrics to evaluate personalization effectively, tackling the cold-start problem in adapting models to sparse user data and addressing stereotypes and biases that may arise in personalized outputs. Privacy concerns surrounding user-specific data are also explored, particularly in balancing personalization with privacy protection. Additionally, we discuss the unique complexities of expanding personalization to multi-modal systems, where integrating user preferences across diverse input types remains an open challenge.

In the remainder of the article, we first present a unifying view and taxonomy for the usage of personalized LLMs (Section 2), and then delve into the theoretical foundations of personalized LLMs (Section 3). Next, we explore the granularity of personalization in LLMs (Section 4), and provide a comprehensive survey and taxonomy of techniques for personalized LLMs (Section 5). We then categorize metrics and methods for the evaluation of personalized LLMs (Section 6), and offer a detailed taxonomy of datasets used for personalized LLMs (Section 7). We discuss the various applications of personalized LLMs (Section 8), and finally, identify key challenges and propose future research directions (Section 9).

2 Unifying Personalized LLMs

To bridge the gap between the two distinct lines of work in the literature, we propose an intuitive taxonomy (see Figure 1) that categorizes personalized LLM efforts into two main categories: ❶ *personalized text generation*, and ❷ *downstream task personalization*. In the first category, *personalized text generation*, the goal is to generate text that directly aligns with individual or group preferences (Salemi et al., 2023; Kumar et al., 2024). For example, a personalized mental health chatbot should generate empathetic responses based on a user's previous conversations, adapting the tone and language to reflect their emotional state. The focus is on producing personalized content, which is evaluated by assessing the quality of the generated text itself using user-written text if available, or alternatively, user preferences and separate reward models that reflect user judgments, as the generated text should match or approximate the style or content the user would produce. In the second category, *downstream task personalization*, personalized LLMs are used to enhance the performance of a specific task, such as recommendation (Lyu et al., 2023; Bao et al., 2023). For

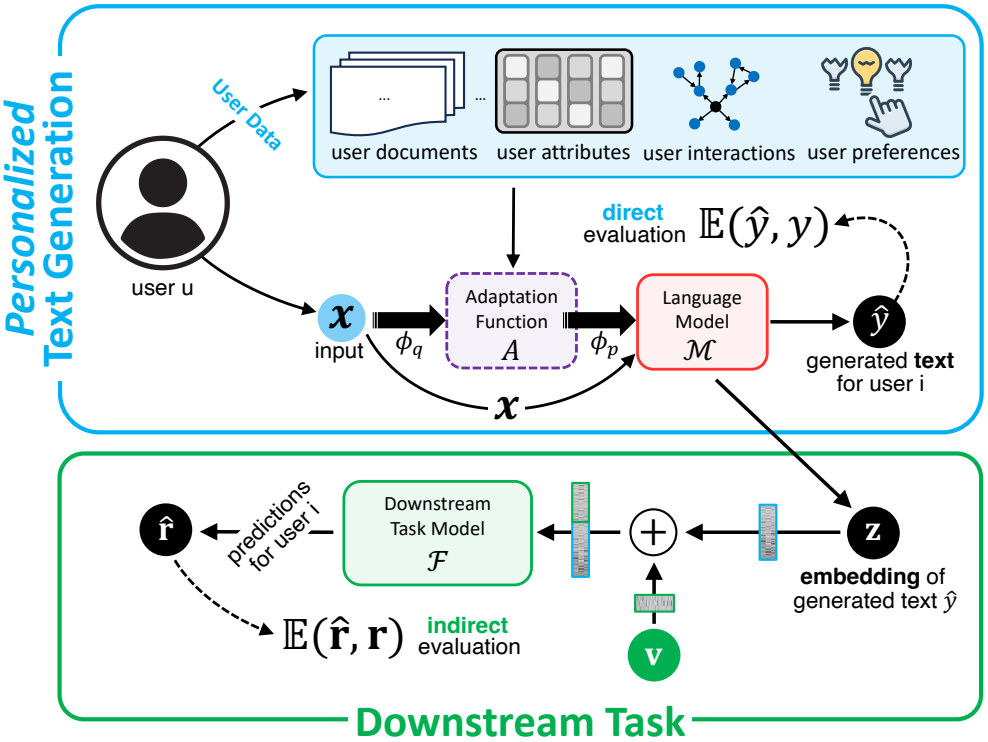


Figure 1: **Taxonomy for Personalized LLM Usage.** To bridge the gap in the existing literature on personalized LLMs, we propose the intuitive taxonomy outlined above, which categorizes work into two main areas. The first focuses on studying the ❶ **personalized text generated** directly, while the second emphasizes using personalized information as intermediate steps or implicitly as embeddings to improve the quality of a ❷ **downstream task** such as recommendation systems. See Section 2 for a detailed discussion. An example of an adaptation function A is a retrieval module. Please note that y here can represent user-written text if available, or alternatively, user preferences and separate reward models that reflect user judgments.

instance, an LLM-enhanced movie recommendation system might suggest new films by analyzing a user's viewing history, preferences, and interactions with previous recommendations. In this scenario, the LLM may generate intermediate tokens or embeddings that enhance the system's performance on specific downstream tasks. While these intermediate tokens are not evaluated directly, they serve as crucial steps in improving the overall effectiveness of the task-specific system. The performance is assessed through task-specific metrics like recommendation accuracy or task success rates. Unlike the first category, this line of work focuses on improving task outcomes rather than the text generation process.

Direct Personalized Text Generation: While there are various techniques for generating personalized text from LLMs, We introduce a general adaptation function A that integrates user-specific information into personalized text generation. An example of A is a retrieval module, with additional concrete examples provided in Figure 7. For a given user u, this information may include user-written documents, static attributes, interaction histories, or preferences, as illustrated in Figure 1. Given a user's textual input x, a query generation function ϕ_q transforms this input to effectively integrate relevant user-specific data via the adaptation function A. The adaptation function A leverages the transformed query $\phi_q(x)$, user data $\mathcal{D}_u$, and optionally a parameter k for flexibility. Subsequently, another transformation function, the personalized prompt generation function ϕ_p, combines the original input x and the output of the adaptation function to form a personalized input $\bar{x}$. Ultimately, the personalized text $\hat{y}$ generated by an LLM $\mathcal{M}$ is:

$$\hat{y} = \mathcal{M}(\phi_p(x, \mathcal{A}(\phi_q(x), \mathcal{D}_u, k))) = \mathcal{M}(\bar{x}) \tag{1}$$

Some existing approaches focus exclusively on generating this personalized text $\hat{y}$ and then directly evaluating its quality. Evaluation typically involves comparing $\hat{y}$ to some form of ground truth or user-specific reference, denoted as y. The reference y may represent actual user-written text when available, or more generally, it may encompass user preferences or evaluations obtained from dedicated reward models or user judgments. The evaluation is performed using a metric $\mathbb{E}(\hat{y}, y)$, such as ROUGE-1, ROUGE-L, METEOR, or other specialized metrics designed for personalized text generation tasks.

While evaluating how well the personalized text generated, $\hat{y}$, matches the known user preferences is crucial, it remains particularly challenging due to the scarcity of datasets with high-quality, user-written labels. This likely contributes to the limited focus on the foundational task of personalized text generation in the literature. Instead, there have been many more works that focus on utilizing the personalized text generated $\hat{y}$ *indirectly* to improve a downstream task such as recommendation or prediction in general. It is important to note that in these indirect approaches, where the focus is on improving a downstream task, the personalized text output $\hat{y}$ is typically not evaluated and is considered less critical. The key consideration is that the generated intermediate text or its embedding when applied to a downstream task, can enhance the overall system's performance. While this line of work lacks interpretability regarding the intermediate information generated by LLMs, it has been demonstrated that augmenting systems with this information typically improves performance in downstream personalization-related applications. For a detailed discussion of the evaluation specific to personalized LLMs, see Sec. 6.

Indirect Downstream Task: Instead of studying how to directly generate text $\hat{y}$ generated for user i, many works focus on leveraging $\hat{y}$ or its personalized embedding $\mathbf{z}$ to improve downstream tasks, such as recommendation. Figure 1 provides an intuitive overview of the fundamental steps used in these approaches. Typically, these methods utilize the embedding $\mathbf{z}$ or $\hat{y}$ as additional information and augment it with other information relevant to the downstream task. In Figure 1, the user-specific embedding $\mathbf{z}$ or intermediate text $\hat{y}$ is augmented with another embedding or task-specific text $\mathbf{v}$ (e.g., concatenated or combined using a function) to form a unified representation that is then passed into the downstream task model $\mathcal{F}$, which can represent any model for a specific application, such as a recommendation system. Although Figure 1 shows a single embedding or text $\mathbf{v}$ combined with $\mathbf{z}$ or $\hat{y}$, in practice, multiple ones or hierarchical combinations can be applied. The downstream model $\mathcal{F}$ then produces predictions $\hat{\mathbf{r}}$, which could include inferred ratings or scores, among other outputs.

While direct ❶ *personalized text generation* and ❷ *downstream task personalization* might appear distinct, they share many underlying components and mechanisms. Both settings often involve retrieving and utilizing user-specific data, constructing personalized prompts or embeddings, and leveraging these to enhance model outputs. The key distinction lies in the dataset they use and the evaluation methods: direct text generation focuses on aligning the generated text with user-written ground-truth, while downstream task personalization evaluates the improvement in specific tasks. Despite these differences, the two approaches can complement each other. For instance, advancements in direct personalized text generation can provide richer, more nuanced intermediate text or embeddings that may enhance downstream tasks. Conversely, improvements in downstream task personalization models can inform better methods for retrieving and leveraging user-specific data in direct generation tasks. By viewing both approaches as two sides of the same coin, researchers from these communities can benefit from cross-pollination. This unification offers an opportunity to share best practices, datasets, and techniques across the two lines of work, driving progress in both areas. In the next section, we delve into these shared foundations, laying out the core principles and formal definitions that unify both lines of work. By framing personalization in a comprehensive theoretical context, we aim to establish a shared vocabulary and methodology that can facilitate cross-disciplinary collaboration between these communities, fostering new insights and innovations in personalized LLMs.

3 Foundations of Personalized LLMs

While previous research (Yang & Flek, 2021; Chen et al., 2024c;b) has explored definitions and analyzed various aspects of personalized LLMs, a comprehensive theoretical framework for understanding and formalizing personalization in these models is still lacking. In this section, we aim to fill this gap by establishing the foundational principles, definitions, and formal structures to formalize the problem of personalization in LLMs.

We systematically develop the necessary notation and conceptual framework to formalize the problem and evaluation, setting the stage for a deeper understanding of how personalization can be effectively implemented and analyzed within LLMs. The following subsections are structured as follows:

- **§3.1 General Principles of LLMs:** We begin by outlining the core principles that form the foundation of LLMs. This provides essential context for understanding how these models function and the underlying mechanics that drive their capabilities.
- **§3.2 Definition of Personalization in LLMs:** We define the term "personalization" within the specific context of LLMs, establishing a clear understanding for subsequent discussions.
- **§3.3 Overview of Personalization Data:** We provide an overview of the current data utilized for personalization, emphasizing the different formats of data sources.
- **§3.4 Formalization of Personalized Generation:** We formalize the conceptual space for personalized generation, providing a structured framework for understanding how personalization can be achieved.
- **§3.5 Taxonomy of Personalization Criteria:** We introduce a comprehensive taxonomy of personalization criteria, categorizing the various factors that influence personalized outputs.

3.1 Preliminaries

Let $\mathcal{M}$ be an LLM parameterized by θ, which takes a text sequence $X \in \mathbb{X}$ as input and produces an output sequence $\hat{Y} \in \hat{\mathbb{Y}}$, where $\hat{Y} = \mathcal{M}(X; \theta)$. The form of $\hat{Y}$ depends on the specific task, with $\hat{\mathbb{Y}}$ representing the output space of possible generations. The inputs can be drawn from a labeled dataset $\mathcal{D} = (X^{(1)}, Y^{(1)}), \cdots, (X^{(N)}, Y^{(N)})$, or from an unlabeled dataset of prompts for sentence continuations or completions $\mathcal{D} = X^{(1)}, \cdots, X^{(N)}$. For this and other notation, see Table 2.

Definition 1 (Large Language Model)**.** *A* large language model (LLM) $\mathcal{M}$*, parameterized by* θ*, is a multi-layer Transformer model with billions (or more) of parameters. It can be structured with an encoder-only, decoder-only, or encoder-decoder architecture and is trained on extensive corpora comprising a vast number of natural language tokens (Zhao et al., 2023; Gallegos et al., 2024).*

Definition 2 (Downstream Tasks)**.** *A* downstream task *is a specific practical application or goal, such as classification, translation, recommendation, or information retrieval, that uses outputs generated by a model (e.g., an LLM). Formally, for a downstream task, we define a corresponding downstream model or function* $\mathcal{F}$ *that takes as input the model's output* $\hat{y}$ *(generated from an initial input* X*) and produces a final result or prediction*

$$\hat{r} = \mathcal{F}(\hat{y})$$

Currently, LLMs are mainly built upon multi-layer Transformer (Vaswani et al., 2017), which employ stacked multi-head attention layers within a deeply structured neural network (Zhao et al., 2023). Based on the use of different components of the original transformer architecture, LLMs can be categorized into the following three categories: (1) decoder-only models (e.g., GPT series (Radford et al., 2018; 2019; Brown et al., 2020; Achiam et al., 2023)) (2) encoder-only models (e.g., BERT-based models (Devlin et al., 2018; Liu et al., 2019)), (3) encoder-decoder models (e.g., T5 (Raffel et al., 2020)). Among those categories, decoder-only LLMs become the most popular type which is optimized for next-token generations.

After pre-training with large-scale unlabeled corpora in an unsupervised manner, the resulting in-context-aware word representations are very effective as general-purpose semantic features for a wide range of NLP tasks. With the scaling of their size and techniques such as instruction tuning (Ouyang et al., 2022; Zhang et al., 2023c; Longpre et al., 2023; Zhou et al., 2024a) and RLHF (Christiano et al., 2017; Stiennon et al., 2020b; Rafailov et al., 2024), LLMs exhibit many emergent abilities (Wei et al., 2022a). This enables LLMs to solve complex tasks and engage in natural conversations with humans, even in a zero-shot manner through text prompting for a wide range of downstream tasks such as sequence classification, text generation, and recommendation (Qin et al., 2023). To further enhance LLMs' performance on specific downstream tasks, models are often fine-tuned with a relatively small amount of task-specific data following the "pre-train, then fine-tune" paradigm, which generally adapts LLMs to particular tasks and achieves better results (Bommasani et al., 2021; Min et al., 2023; Liu et al., 2023b).

Definition 3 (Prompt)**.** *A* Prompt $\mathcal{H}$ *is a specific input or set of instructions provided to a language model, which guides its generation of text* $M(X;\theta)$*. Prompts can vary in complexity from simple word or phrase completions to detailed, structured contexts or questions aimed at eliciting specific types of responses or performing certain tasks. Prompts can be multi-modal, including text, image, audio, or video inputs.*

- ***System Prompt:*** *A* System Prompt $\mathcal{H}_{sys}$ *is a predefined prompt that initializes the interaction, setting the overall behavior, style, or constraints of the language model. It often provides consistent instructions on how the model should respond to subsequent user prompts throughout the interaction. This is particularly useful for role-playing or establishing the model's tone and persona.*
- ***User Prompt:*** *A* User Prompt $\mathcal{H}_{usr}$ *is an input provided by the user during the interaction with the language model, typically seeking specific information, responses, or actions from the model. For simplicity, in the following sections, we will represent the user prompt as* x*.*

3.2 Formulation of Personalization

Definition 4 (Personalization)**.** Personalization *refers to the process of tailoring a system's output to meet the individual preferences, needs, and characteristics of an individual user or a group of users. In the context of LLMs, personalization involves adjusting the model's responses based on user-specific data, historical interactions, and contextual information to enhance user satisfaction and relevance of the entire system's generated content.*

Definition 5 (User Preferences)**.** User Preferences *refer to the specific likes, dislikes, interests, and priorities of an individual user or a group of users. These preferences guide the personalization process by informing the system about the desired characteristics and features of the output. In the context of LLMs, user preferences can be derived from explicit feedback (e.g., pairwise comparison), historical interactions, and contextual signals to tailor responses and improve the relevance and satisfaction of the generated content.*

Definition 6 (Personalized Large Language Model)**.** *A* Personalized Large Language Model *(Personalized LLM)* $\mathcal{M}_p$ *is an LLM that has been adapted to align with the individual preferences, needs, and characteristics of a specific user or group of users. This adaptation involves utilizing user-specific data, historical interactions, and contextual information to modify the model's responses, making them more relevant and satisfying for the user. Personalized LLMs aim to enhance the user experience by providing tailored content that meets the unique expectations and requirements of the user.*

Definition 7 (User Documents)**.** User Documents $\mathcal{D}_u$ *refer to the collection of texts and writings generated by a user* u*. This includes reviews, comments, social media posts, and other forms of written content that provide insights into the user's preferences, opinions, and sentiments.*

Definition 8 (User Attributes)**.** User Attributes $A_u = \{a_1, a_2, \ldots, a_k\}$ *are the static characteristics and demographic information associated with a user* $u \in U$*. These attributes include age, gender, location, occupation, and other metadata that remain relatively constant over time.*

Definition 9 (User Interactions)**.** User Interactions $I_u = \{i_1, i_2, \ldots, i_m\}$ *capture the dynamic behaviors and activities of a user* $u \in U$ *within a system. This includes actions such as clicks, views, purchases, and other engagement data that reflect the user's preferences and interests.*

Personalization, the practice of tailoring experiences to the preferences of individual users or groups of users, is crucial for bridging the gap between humans and machines (Rossi et al., 1996; Montgomery et al., 2004; Chen et al., 2024c). Such experiences can include aligning with specific user or group preferences, adjusting the style or tone of generated content, and providing recommendation items based on the user's interaction history in a wide range of downstream tasks. Users can be actual individuals with a history of interactions or described by specific characteristics such as demographic information, allowing both humans and machines to better understand and cater to their needs. In this work, instead of just focusing on personalization for single individual users, we aim to formalize and clarify the term "personalization" by categorizing its objectives based on the size of the targeted group. We classify personalization into three categories based on their focus: aligning with the preferences of individual users, groups of users, or the general public (Sec. 4). Additionally, these three levels of personalization enable the incorporation of different types of input data, each contributing uniquely to the personalization process. It is important to note that not all fine-tuning

equates to personalization. For example, most supervised fine-tuning practice is a process where models are trained on specific datasets to perform better on a downstream task. However, only fine-tuning that adjusts a model to cater to specific user or group preferences—such as adapting a model to a user's writing style or content preferences—counts as personalization. In contrast, fine-tuning on a general corpus to improve overall task performance is not personalized, as it does not address the unique preferences of individuals or groups. This distinction is key to understanding the objectives of personalized LLMs across the different levels of granularity.

3.3 Personalization Data

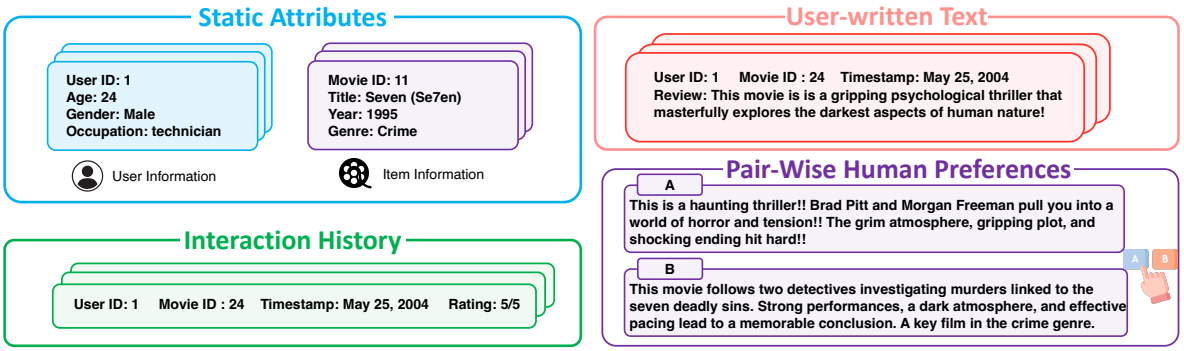


Figure 2: **Overview of Personalization Data.** This figure presents an overview of the various types of user-specific data used in downstream personalization tasks. It categorizes the data into three primary formats: (i) **Static Attributes**, which include demographic information and item metadata that remain relatively constant over time; (ii) **Interaction History**, capturing dynamic user behaviors and preferences through previous activities and engagement data; (iii) **User-Written Text**, encompassing reviews, dialogues, and social media posts that provide rich insights into user sentiment and preferences; and (iv) **Pair-Wise Human Preferences**, explicit feedback or annotations that guide the system to align with individual user needs.

In this section, we provide an overview of various formats of user-specific information commonly used in downstream personalization tasks. Understanding such data is critical for leveraging user information and designing targeted personalization techniques to enhance the performance of LLMs in diverse applications. Figure 2 illustrates this overview with concrete examples.

3.3.1 Static Attributes

Static attributes refer to information about both users and items that remain relatively constant over time. These attributes form the foundation of many personalization strategies and are often used to segment users and items for more targeted recommendations. Except for unique identifiers assigned to each user and item, such as User ID and Item ID, common static attributes include:

- **User's Demographic Information**: Age, gender, location, and occupation can help infer preferences and tailor content or product recommendations.
- **Item Information**: For recommendation systems, item-specific data, such as title, release date, genre, and other relevant metadata, play a crucial role in understanding user preferences and making accurate recommendations.

Static attributes provide a reliable basis for long-term personalization strategies. Typically collected during user registration or profile setup for users, and during the cataloging process for items, this data requires minimal human effort for annotation. However, static attributes do not capture changes in user preferences or item relevance over time, which limits their effectiveness in downstream personalization tasks. Additionally, collecting and storing demographic information can raise privacy issues, necessitating careful handling and

compliance with data protection regulations. Techniques for anonymizing data (Samarati & Sweeney, 1998) are essential to address these concerns.

3.3.2 Interaction History

Interaction history captures the dynamic aspects of a user's behavior and preferences based on their interactions with a system. This data is crucial for understanding user preferences and enabling real-time personalized recommendations. Interaction history includes information about past activities, such as movies watched, songs listened to, items purchased, or articles read. It also covers user interactions with items they have clicked on or viewed, including engagement duration, which helps infer interests and engagement levels. Additionally, in the context of interactions with LLMs, this history includes the content of previous prompts, responses, and the patterns of user engagement with the generated outputs, all of which contribute to tailoring future interactions.

The advantage of interaction history is its dynamic and up-to-date nature, providing real-time insights into user preferences and enabling timely and relevant recommendations. Detailed interaction data offers rich context, aiding in a deeper understanding of user behavior. However, interaction history can be voluminous and complex to process, requiring sophisticated data-handling techniques. Additionally, past interactions may not always accurately reflect current preferences, necessitating careful analysis to maintain relevance.

3.3.3 User-Written Text

User-written text includes any form of written content generated by users, such as reviews, comments, dialogues, or social media posts. This type of data is rich in user sentiment and can provide deep insights into user preferences and opinions. User text data typically encompasses:

- **Reviews**: Written evaluations of products or services, often including ratings and detailed comments. For example, the Amazon Review Data (Ni et al., 2019) contains 233.1 million reviews, offering insights into user experiences and preferences through detailed textual feedback and ratings.
- **Dialogues and Conversations**: Textual exchanges between users and dialogue systems or other users. The ConvAI2 (Dinan et al., 2020) dataset includes dialogues where participants are assigned personas and engage in natural conversations, helping to understand user interaction patterns and improve conversational agents.
- **Social Media Posts**: Short messages or comments on platforms like Reddit, Twitter, or Facebook, which can be analyzed to understand user sentiments and trends.

In the context of LLMs, this also includes human-written exemplars often used for few-shot learning, reflecting user preferences or intent to guide the model's responses. The potential use cases for user text data are extensive. For example, sentiment analysis (Medhat et al., 2014; Wankhade et al., 2022) can be performed to understand user opinions and improve product offerings or customer service. Conversational agents can be enhanced by analyzing user conversations to make interactions more natural and engaging. The advantages of user text data lie in its depth of insight, providing detailed information about user preferences, opinions, and sentiments. It is versatile and applicable across various domains, from product reviews to social media analysis. However, text data is inherently unstructured, necessitating advanced NLP techniques for effective analysis. Besides, comprehensively evaluating such nuanced data, especially for personalization, is challenging with existing metrics. Additionally, user-generated content can be noisy and variable in quality, complicating accurate analysis. Annotating high-quality new data points is expensive, further adding to the complexity.

3.3.4 Pair-Wise Human Preferences

Pair-wise human preferences refer to explicit user feedback indicating their preferred responses from a set of candidate outputs. This data format typically involves human annotations selecting the most desired option, making it essential for training models to align closely with individual user needs and preferences. Unlike static attributes or interaction history, pair-wise preferences offer highly specific and direct feedback, serving as explicit instructions on how users expect the model to behave or respond in given scenarios. For example, users might specify whether they want a response to be easily understood by a layperson or tailored for an

expert. In this way, users can explicitly state what they want, reducing ambiguity and implicitly, which can be useful leading to higher user satisfaction and more effective personalization. However, designing an appropriate alignment strategy remains a significant challenge for personalization applications. Most current works focus on aligning models with general, aggregate human preferences, rather than diverse, individual perspectives (Jang et al., 2023). Developing methods to capture and use these individual direct preferences effectively is essential for advancing personalized systems.

Definition 10 (Alignment)**.** Alignment $\mathcal{G}$ *is the process or state by which an AI system's goals,* $\mathcal{G}_A$*, is consistent with human values and intentions, denoted as* $\mathcal{G}_H$*. Mathematically, alignment can be defined as ensuring that the behavior policy* π_A *of the AI system maximizes the utility function* U_H *representing human values. Formally,*

$$\mathcal{G} = \{\pi_A \mid \pi_A \in \arg\max_{\pi} \mathbb{E}_{\pi}\left[U_H\right]\}$$

where π_A *is the policy of the AI system,* $\mathbb{E}_{\pi}\left[U_H\right]$ *is the expected utility under policy* π*, and* $\arg\max_{\pi}$ *denotes the set of policies that maximize the expected human utility* U_H*.*

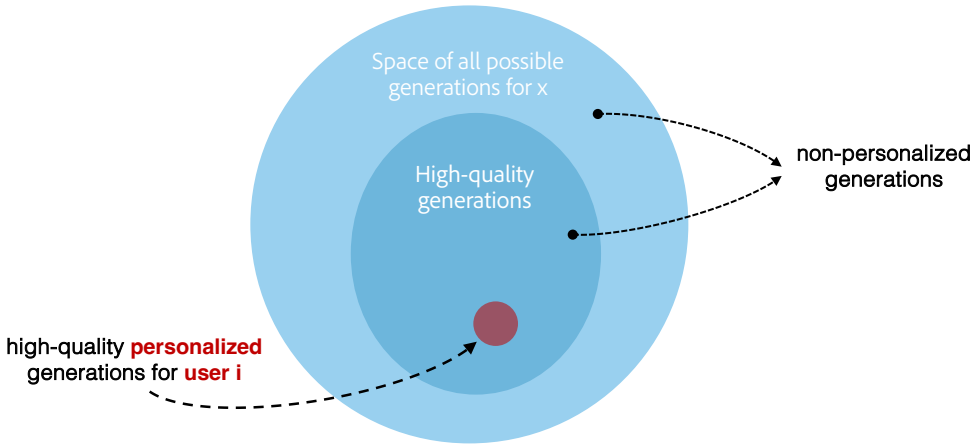


Figure 3: **Space of Personalized Generations.** We characterize the space of generations for query x, including the space of all possible generations $\mathcal{S}(x)$, the space of all high quality generations $\mathcal{S}_h(x)$, and finally, the space of user-specific high-quality personalized generations $\mathcal{S}_i(x)$ for user i. Intuitively, given two users i and j, the space of high-quality personalized generations for each user may be completely disjoint.

3.4 Space of Personalized Generations

In this section, we briefly formalize and analyze the problem of personalized LLMs and their solution spaces and provide an intuitive overview in Figure 3. This serves two purposes: providing intuition on the difficulty of the problem and characterizing the properties and unique advantages that relate to other well-studied problems.

Let us first establish the formalization of the personalized LLM problem. Consider a generic input example $x \in \mathcal{X}$. We denote the generative model by $g : \mathcal{Z} \times \mathcal{X} \to \mathcal{Y}$, where $\mathcal{Z}$ represents the latent space and $\mathcal{Y}$ represents the space of all possible generations. Given input $x \in \mathcal{X}$, the space of all possible generations is

$$\mathcal{S}(x) = \{g(\mathbf{z}, x) : \mathbf{z} \in \mathcal{Z}\} \subseteq \mathcal{Y}$$

To facilitate a comprehensive understanding of personalization, we delineate the following sets:

- The space of all possible generations $\mathcal{Y}$.
- The space of all possible generations for a given input x is $\mathcal{S}(x)$.
- The space of high-probability generations for a given input x, denoted by:

$$\mathcal{S}_h(x) = \{y \in \mathcal{Y} : P(y|x) \geq \delta\}$$

where $P(y|x)$ is the probability of generation y given input x and δ is a threshold representing high-quality content.

- The space of user-specific generations for a user $u_i \in U$ given input x, denoted by:

$$\mathcal{S}_i(x) = \{y \in \mathcal{S}_h(x) : f(P_{u_i}, y) \leq \epsilon, P(y|x) \geq \delta\}$$

where $f(P_{u_i}, y)$ is a function that quantifies the alignment of the generation y with the user's preferences P_{u_i}, and ϵ is a threshold for user-specific relevance.

An intuitive overview of the space of personalized generations for a specific user can be found in Figure 3. Note that the space of user-specific generations $\mathcal{S}_i(x)$ is significantly smaller and more targeted compared to the space of all possible generations $\mathcal{S}(x)$. One example is that there may be many correct responses to a specific question, however, only a very small subset of answers may capture the important details needed for the answer to be useful to the specific user. In particular, the user may need additional steps to carry out a specific task, or they may know certain terminology and language that would enable better understanding for the user. Overall, the key takeaways are that the user personalization tasks are extremely challenging, though will only become increasingly important in the future. In particular, there is only a very tiny space of responses that are useful to a specific user, and generating such a response is only becoming more and more important. This highlights the need to not only develop better techniques to generate such user-specific responses, but also better data and evaluation metrics to quantify it.

3.5 Personalization Criterion Taxonomy

When evaluating the personalization of generated text in LLMs, it is essential to consider several critical aspects to ensure the content is effectively tailored to individual users. These aspects constitute a taxonomy of personalization criteria, encompassing various dimensions of personalized content generation.

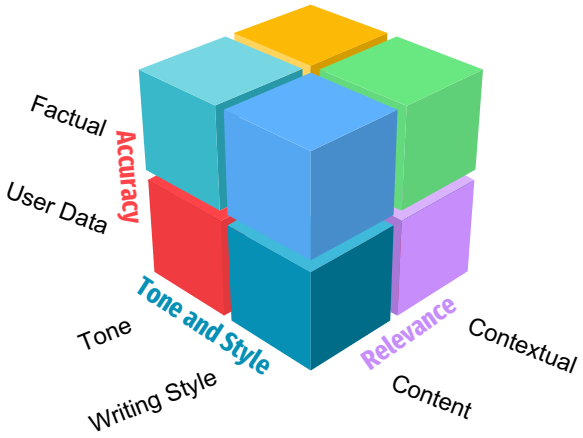


Figure 4: **Dimensions in Personalized Criterion.** We propose a framework that expands the dimensions of personalization criteria LLMs along three aspects: (i) Tone and Style, which includes writing style and tone preferences to match the user-written text; (ii) Relevance, encompassing content relevance to user interests and contextual relevance for specific situations; and (iii) Accuracy, which ensures both factual correctness and accurate representation of user data. These aspects interact to form a comprehensive taxonomy, addressing the multi-faceted nature of effective personalization in LLM-generated text.

Tone and Style One of the fundamental aspects of personalized text generation is the alignment of tone and style with the user's preferences and previous interactions. This includes:

- **Writing Style**: The writing style should be consistent with the user's preferred style or previous interactions. For instance, if a user typically prefers a more concise style for an email, the generated text should reflect such preference, ensuring a seamless user experience.

- **Tone**: The tone of the generated content should match the user's preferred tone, which could vary depending on the context. For example, the tone could be formal, casual, professional, or friendly, depending on the user's past written texts and the situational requirements.

Relevance Personalization also necessitates that the generated content be highly relevant to the user's interests, preferences, and current needs. This relevance is assessed on two levels:

- **Content Relevance**: This criterion evaluates whether the content aligns with the user's interests and preferences. It ensures that the generated text is pertinent and valuable to the user, thus enhancing engagement and satisfaction. For example, if a user has recently shown interest in sustainability topics, the LLM should prioritize generating content related to green technologies or eco-friendly practices in relevant contexts, such as when drafting blog posts or social media updates.
- **Contextual Relevance**: Beyond general interests, it is crucial that the content is appropriate for the specific context or situation in which the user will encounter it. For example, if the user is preparing for a business presentation, the LLM should focus on generating content that is formal, data-driven, and aligned with the specific industry, rather than casual or unrelated topics.

Accuracy Accuracy is another critical dimension of personalized text generation, ensuring that the information provided is reliable and precise. This includes:

- **Factual Accuracy**: The generated content should be factually correct and based on reliable information. This ensures the credibility of the content and maintains the trust of the user. For example, if the LLM is generating a report on recent market trends, it should use up-to-date data and cite reliable sources, avoiding any outdated or incorrect information.
- **User Data Accuracy**: Personalization heavily depends on the accuracy of the user data used to tailor the content. The personalized content must be based on up-to-date and correct user data, which includes the user's preferences, past behavior, and interactions. For example, if a user recently changed their job title from 'Manager' to 'Director,' the LLM should generate emails or documents that reflect this new role and associated responsibilities, rather than using outdated information.

These aspects of personalization—tone and style, relevance, and accuracy—form the foundation of a robust taxonomy for evaluating personalized LLMs. Each criterion plays a vital role in ensuring that the generated content is tailored effectively, providing a unique and satisfying user experience. This taxonomy not only aids in the systematic evaluation of personalized LLMs but also highlights the multi-faceted nature of personalization. By addressing each of these criteria, researchers and practitioners can develop more sophisticated and user-centric language models that better serve the diverse needs and preferences of users.

Table 1 provides an illustrative breakdown of these criteria, along with their respective descriptions and examples.

Table 1: **Taxonomy of Personalized LLM Criterion.**

Criterion	Description and Examples
Tone and Style	
Writing Style	Is the writing style consistent with the user's preferred style or previous interactions?
Tone	Does the tone of the text match the user's preferences (previous written text) and context (*e.g.*, formal, casual, etc)?
Relevance	
Content Relevance	Does the content match the user's interests, preferences, and needs?
Contextual Relevance	Is the content appropriate for the specific context/situation that the user will encounter it?
Accuracy	
Factual Accuracy	Are the facts and information presented in the text correct and reliable?
User Data Accuracy	Is the personalized content based on accurate and up-to-date user data?

3.6 Overview of Taxonomies

In this section, we present a high-level summary of each taxonomy proposed in the subsequent sections of the paper. Comprehensive descriptions of these taxonomies can be found in Sections 4, 5, 6, and 7.

3.6.1 Taxonomy of Personalization Granularity of LLMs

We propose three different levels of personalization granularity for LLMs, each addressing different scopes of personalization. These levels help in understanding the depth and breadth of personalization that can be achieved with LLMs. The three levels are:

- §4.1 **User-level Personalization:** Focuses on the unique preferences and behaviors of a single user. Personalization at this level utilizes detailed information about the user, including their historical interactions, preferences, and behaviors, often identified through a user ID.
- §4.2 **Persona-level Personalization:** Targets groups of users who share similar characteristics or preferences, known as personas. Personalization here is based on the collective attributes of these groups, such as expertise, informativeness, and style preferences.
- §4.3 **Global Preference Personalization:** Encompasses general preferences and norms that are widely accepted by the general public, such as cultural standards and social norms.

3.6.2 Taxonomy of Personalization Techniques for LLMs

We categorize personalization techniques for LLMs based on the way user information is utilized. These techniques provide various methods to incorporate user-specific data into LLMs to achieve personalization. The main categories are:

- §5.1 **Personalization via Retrieval-Augmented Generation:** Incorporates user information as an external knowledge base, encoded through vectors, and retrieves relevant information using embedding space similarity search for downstream personalization tasks.
- §5.2 **Personalization via Prompting:** Incorporates user information as the context within the prompts for LLMs, allowing for downstream personalization tasks.
- §5.3 **Personalization via Representation Learning:** Encodes user information into the embedding spaces of neural network modules, which can be represented through model parameters or explicit embedding vectors specific to each user.
- §5.4 **Personalization via Reinforcement Learning From Human Feedback:** Uses user information as the reward signal to align LLMs with personalized preferences through reinforcement learning.

3.6.3 Taxonomy of Evaluation Methodologies for Personalized LLMs

Evaluation metrics for personalized LLMs can be classified based on how they measure the effectiveness of personalization. These metrics ensure that the personalized outputs meet the desired standards of relevance and quality. The main categories are:

- §6.1 **Intrinsic Evaluation:** Evaluates the personalized text generated directly, focusing on factors like personalized content, writing style, and more.
- §6.2 **Extrinsic Evaluation:** Relies on downstream applications such as recommendation systems to demonstrate the utility of the generated text from the personalized LLM.

3.6.4 Taxonomy of Datasets for Personalized LLMs

We propose a taxonomy that categorizes personalized LLM datasets based on whether they contain text written by specific users. This helps in understanding the data's role in training or evaluating personalized LLMs directly or indirectly. The main categories are:

§7.1 Personalized Datasets *with* Ground-Truth Text: Contain actual ground-truth text written by users, enabling direct evaluation of personalized text generation.

§7.2 Personalized Datasets *without* Ground-Truth Text: Used for indirect evaluation via downstream applications, as they do not contain user-specific ground-truth text.

Table 2: **Summary of key notation.**

Notation	Definition
$\mathcal{D}$	dataset
$\mathcal{D}_i$	user i's specific user data
t_i	text written by user i
a_i	attributes/preferences of user i
I_i	interactions of user i
$X_i = (x_1, \cdots, x_m) \in \mathbb{X}$	generic input text for user i
$\bar{x}$	transformed personalized input based on retrieved user information
$Y_i \in \mathbb{Y}$	ground-truth text for user i
$\hat{Y}_i \in \hat{\mathbb{Y}}$	personalized text generated for user i
$\mathbf{v}$	task-specific information
U	a set of users
i	a single user $i \in U$
$\mathcal{S}$	set of personas
S	a persona $S \in \mathcal{S}$
P_u	the set of a single user's preferences
P_s	the set of a persona s's preferences
P_G	global preferences
$\mathbf{r}$	downstream tasks' label
$\hat{\mathbf{r}}$	model's predictions on downstream task
$\mathcal{G}$	the process of alignment
$\mathcal{G}_{\mathcal{A}}$	AI system's targeted preferences
$\mathcal{G}_{\mathcal{H}}$	human's values and intentions
π_A	AI system's behavior policy
U_H	the utility function which represents human values
$\mathbb{E}_i$	intrinsic evaluation
$\mathbb{E}_e$	extrinsic evaluation
$\psi(\cdot) \in \Psi$	an evaluation metric
$\psi_a(\cdot) \in \Psi_a$	an evaluation metric for downstream application
$\mathcal{L}(\cdot)$	loss function
$\mathcal{M}$	LLM
$\mathcal{M}_p$	personalized LLM
$\mathcal{H}_{\text{sys}}$	the system prompt to input LLMs
$\mathcal{H}_{\text{usr}}$	the user prompt to input LLMs
g	generative model
$Emb(\cdot)$	word or sentence encoder, which can be a part of $\mathcal{M}$
$\mathcal{R}(\cdot)$	retrieval model
$A(\cdot)$	adaptation function that incorporate user-specific information to personalized LLMs
$RecSys$	recommendation system
$\mathcal{F}(\cdot)$	downstream model such as recommendation system
ϕ_q	query construction function
ϕ_p	personalized prompt construction function
$\mathbf{z}$	embedding generated from user information

4 Personalization Granularity of LLMs

Definition 11 (Personalization Granularity)**.** Personalization Granularity *refers to the level of detail at which personalization objectives are defined and implemented. It determines the extent to which the system's*

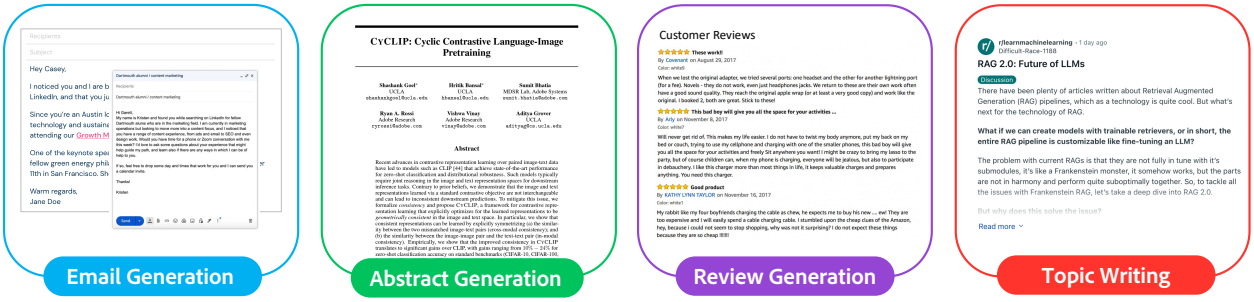
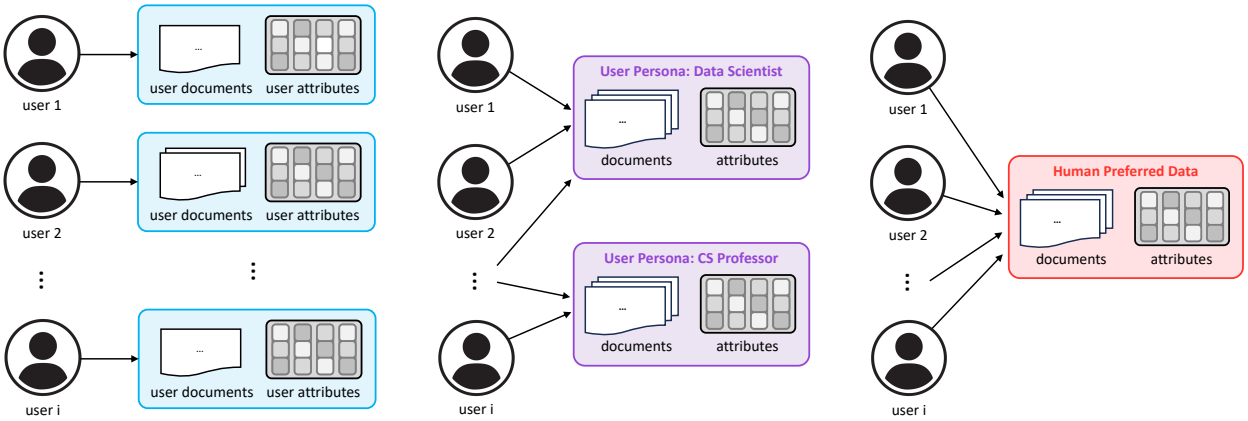


Figure 5: **Examples of Personalization Tasks and Data.**



(a) User-level Personalization (§4.1) (b) Persona-level Personalization (§4.2) (c) Global Preference Personalization (§4.3)

Figure 6: **Personalization Granularity Taxonomy.**

responses are tailored to specific criteria, such as individual users, groups of users with a certain shared persona, or the general public, influencing how finely or broadly the personalization is applied.

In this section, we propose a taxonomy for personalized LLMs based on the granularity of the personalization objective. Specifically, personalized LLMs can be categorized by their focus on aligning with the preferences of individual users, groups of users, or the general public. In this survey, we formally define the granularity of personalization with the following distinctions:

- **User-level Personalization (Sec. 4.1)**: This level focuses on the unique preferences and behaviors of a single user. Personalization at this level utilizes detailed information about the user, including their historical interactions, preferences, and behaviors, often identified through a user ID (Li et al., 2024g). Formally, let U represent the set of users, and $P_u = \{p_u^1, p_u^2, ..., p_u^n\}$ denote the set of personalized preferences for user $u \in U$. The objective function of the downstream task is $\mathcal{L}_{task}$. The objective of personalization on this level is to minimize this function:

$$\theta^* = \underset{\theta}{\operatorname{argmin}}\, \mathcal{L}_{task}(f_\theta(P_u))$$

  where θ can be parameters or prompts in the LLM-based system f.

- **Persona-level Personalization (Sec. 4.2)**: This level targets groups of users who share similar characteristics or preferences, known as personas. Personalization here is based on the collective attributes of these groups, such as expertise, informativeness, and style preferences (Jang et al., 2023). Formally, let S represent the set of personas, where each persona $s \in S$ is composed of a subset of users $U_s \subseteq U$ with shared characteristics or preferences. Let P_g denote the set of personalized

preferences for persona s. For every preference $p_i \in P_s$ and every user $u \in U_s$, it holds that $p_i \in P_u$. The objective of personalization on this level is to minimize this function:

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \mathcal{L}_{task}(f_\theta(P_s))$$

- **Global Preference Personalization (Sec. 4.3)**: This level encompasses general preferences and norms that are widely accepted by the general public. For example, broadly accepted cultural standards and social norms. Formally, let P_{global} represent the set of universal preferences. For every preference $p_i \in P_{global}$ and every user $u \in U$, it holds that $p_i \in P_u$. The objective of personalization on this level is to minimize this function:

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \mathcal{L}_{task}(f_\theta(P_{global}))$$

4.1 User-level Personalization

In this section, we discuss user-level personalization, which focuses on data at the individual level (Zollo et al., 2024). As depicted in Figure 6(a), this type of personalization focuses on optimizing preferences for each user uniquely identified by a user ID. For instance, in the MovieLens-1M recommendation dataset (Harper & Konstan, 2015), each user has demographic information such as *UserID*, *Gender*, *Age*, *Occupation*, and *Zip-code*, alongside corresponding movie interactions (MovieID, Rating, Timestamp). The goal is to recommend new movies based on each user's profile and viewing history. The advantage of this level of personalization is that it offers the most fine-grained approach, minimizing noise from other users. This is particularly beneficial in domains such as online shopping, job recommendations (Wu et al., 2024b), and healthcare (Abbasian et al., 2023; 2024; Zhang et al., 2024a; Jin et al., 2024b), where individual user behavior can vary significantly, and such detailed, individualized personalization is crucial. One of the main challenges of this level of personalization is the "cold-start problem" which refers to users with minimal interaction history, often termed "lurkers" in recommendation systems (Sun et al., 2024). However, many studies (Salemi et al., 2023; Rajput et al., 2023; Xi et al., 2023) choose to remove such data during the preprocessing stages. This exclusion potentially undermines the robustness of the systems by disregarding the subtleties and potential insights offered by these underrepresented user interactions.

4.2 Persona-level Personalization

In this section, we discuss persona-level personalization, where the input comprises the preferences of users categorized by group or persona. As illustrated in Figure 6(b), this approach targets optimizing the preferences of a user group sharing common characteristics. A natural language description encapsulating these shared traits represents the entire group within prompts or other relevant components. For example, Jang et al. (2023) design three distinct perspectives of preferences: expertise, informativeness, and style, with each dimension featuring two conflicting personas or preferences. For instance, in the expertise dimension, one persona prefers content that is easily understandable by an elementary school student, while the other prefers content that is comparable only to a PhD student in the specific field. From this example, we can observe that, compared to localized user-specific personalization (Sec. 4.1), each persona represents a broader portrait of a group of users, focusing on more general features rather than detailed user-specific information. The advantage of persona-level personalization lies in its effectiveness in scenarios where shared characteristics are prominent and crucial for downstream tasks, while user-specific attributes are less significant. Additionally, once these characteristics are extracted, this data format is easier to process, either by including it directly in the prompt or utilizing it through RLHF, compared to lengthy user-specific profiles. However, extracting representative characteristics through natural language descriptions can be challenging in practice, often requiring substantial reliance on human domain knowledge.

4.3 Global Preference Personalization

In many applications, only global user preference data may be available, representing the preferences of the entire population rather than those of individual users. While this falls outside the primary scope of

personalization in this survey, we include a discussion of it for completeness. These preferences typically encompass human values expected to be accepted by the general public, such as social norms, factual correctness, and instruction following (Taylor et al., 2016; Gabriel, 2020; Liu, 2024). The common format of such data includes a given instruction, multiple options, and a label annotated by human annotators indicating which option is preferable (Ethayarajh et al., 2022; Stiennon et al., 2020a; Nakano et al., 2021; Bai et al., 2022; Ganguli et al., 2022). These datasets are typically used through RLHF to align LLMs. The advantage of global preference alignment is its potential to enhance LLMs in terms of safety (Gehman et al., 2020; Ge et al., 2023; Anwar et al., 2024; Ji et al., 2024a), social norms (Ryan et al., 2024), and ethical issues (Liu et al., 2021; Rao et al., 2023), ensuring they align with human values. However, the disadvantage is that it may introduce noise, as individual preferences can vary and may not always represent the general public accurately. Moreover, this level of alignment does not capture fine-grained personalization.

4.4 Discussion

The granularity of personalization in LLMs involves trade-offs between *precision*, *scalability*, and *richness* of personalized experiences. User-level personalization offers high precision and engagement but faces challenges with data sparsity and scalability. Persona-level personalization is efficient and representative but less granular and requires domain knowledge for defining personas. Global preference personalization provides broad applicability and simplicity but lacks specificity and can introduce noise from aggregated data. In the future, hybrid approaches may leverage the strengths of each method while mitigating their weaknesses. For instance, a hierarchical personalization framework can combine user-level personalization for frequent users, persona-level personalization for occasional users, and global preferences for new users. This balances precision and scalability by tailoring experiences based on user interaction levels. Another idea is *context-aware personalization*, which starts with persona-level personalization and transitions to user-level as more data becomes available, addressing the cold-start problem. This approach allows the system to offer relevant personalization initially and gradually refine it with detailed user-specific data. Such adaptive systems can dynamically adjust the granularity based on user engagement, context, and data availability. These systems can switch between levels of personalization, providing a balanced and effective user experience by utilizing the most appropriate granularity for each situation. Integrating information across different granularities may further enhance personalization. User-level data can refine persona definitions, making them more accurate and representative. Conversely, persona-level insights can inform user-level personalization by providing context on shared characteristics. Global preferences can serve as a baseline, ensuring that individual and persona-level personalization aligns with broadly accepted norms and values. Currently, datasets for these three levels of granularity are often orthogonal and unrelated. Developing datasets that encompass user-level, persona-level, and global preferences is crucial. Such datasets would enable more seamless integration and transition between different levels of personalization, enhancing the robustness and effectiveness of LLMs in catering to diverse user needs. In conclusion, the choice of personalization granularity should be guided by specific application requirements, balancing precision, scalability, and the ability to provide rich, personalized experiences. Hybrid approaches and integrated datasets are key to achieving optimal personalization outcomes.

5 Taxonomy of Personalization Techniques for LLMs

In this section, we propose a taxonomy of personalization techniques for LLMs categorized by the way user information is utilized. In particular, techniques for personalizing LLMs are categorized as follows:

- **Personalization via RAG (Sec 5.1)**: This category of methods incorporates user information as an external knowledge base, encoded through vectors. When new inference data arrives, the relevant information is retrieved using embedding space similarity search for downstream personalization tasks.
- **Personalization via Prompting (Sec 5.2)**: This category of methods incorporates user information as the context within the prompts for LLMs. By providing this contextual information, LLMs can either directly perform downstream personalization tasks through text generation or

Table 3: **Taxonomy of Techniques for Personalized LLMs.**

Category	Mechanism
Personalization via RAG (§ 5.1)	Sparse Retrieval (§ 5.1.1)
	Dense Retrieval (§ 5.1.2)
Personalization via Prompting (§ 5.2)	Contextual Prompting (§ 5.2.1)
	Persona-based Prompting (§ 5.2.2)
	Profile-Augmented Prompting (§ 5.2.3)
	Prompt Refinement (§ 5.2.4)
Personalization via Representation Learning (§ 5.3)	Full-Parameter Fine-tuning (§ 5.3.1)
	Parameter-Efficient Fine-tuning (§ 5.3.2)
	Embedding Learning (§ 5.3.3)
Personalization via RLHF (§ 5.4)	

act as intermediate modules to extract more relevant information, thereby enhancing the system's performance on downstream tasks.

- **Personalization via Representation Learning (Sec 5.3)**: This category of methods encodes user information into the embedding spaces of neural network modules. The user information can be represented through the entire parameters of the LLM, a subset of the model's parameters, a small number of additional parameters, or an explicit embedding vector specific to each user.
- **Personalization via RLHF (Sec 5.4)**: This category of methods uses user information as the reward signal to align LLMs with personalized preferences through reinforcement learning.

The following sections describe different techniques for achieving personalization in downstream tasks. Note that most of these approaches are orthogonal to each other, meaning they can coexist within the same system. We provide a summary of the personalization techniques organized intuitively using the proposed taxonomy in Table 3.

5.1 Personalization via Retrieval-Augmented Generation

Retrieval-augmented generation (RAG) enhances LLM performance by retrieving relevant document segments from an external knowledge base using semantic similarity calculations (Gao et al., 2023). This approach is widely employed in information retrieval and recommendation systems (Zhao et al., 2024b; Rajput et al., 2023; Di Palma, 2023; Wang et al., 2024g). While RAG can reduce hallucinations by grounding the generation process in retrieved factual content (Shuster et al., 2021; Li et al., 2024d), these retrieval modules can also be used to retrieve personalized information, enabling the generation of customized, tailored outputs.

Definition 12 (Retrieval Model)**.** *A* Retrieval Model $\mathcal{R}$ *is a system designed to identify and return relevant information from a large external database* $\mathcal{D}$ *in response to a query* $q \in \mathcal{Q}$*. Given a query* q*, the retrieval model aims to find the document or data point* $d^* \in \mathcal{D}$ *that maximizes the relevance function* $r(q, d)$*:*

$$d^* = \arg\max_{d \in \mathcal{D}} r(q, d)$$

where d *represents an individual document or data point in* $\mathcal{D}$*.*

Definition 13 (Retrieval-augmented Generation)**.** Retrieval-augmented Generation *(RAG) is a process in which a language model* $\mathcal{M}$ *leverages a retrieval model* $\mathcal{R}$ *to enhance its generation capabilities. Given an input* X_u *from user* u*, the retrieval model* $\mathcal{R}$ *identifies* k *relevant external data points or documents from a dataset* $\mathcal{D}$*. These retrieved data points are then incorporated into the input text to form a transformed input* $\bar{x}$*, which is used by the language model* $\mathcal{M}$ *to generate a grounded output* $\hat{y}_u$*.*

Formally, the process can be described as follows:

$$\bar{x} = \phi_p(X_u, \mathcal{R}(\phi_q(X_u), \mathcal{D}, k))$$

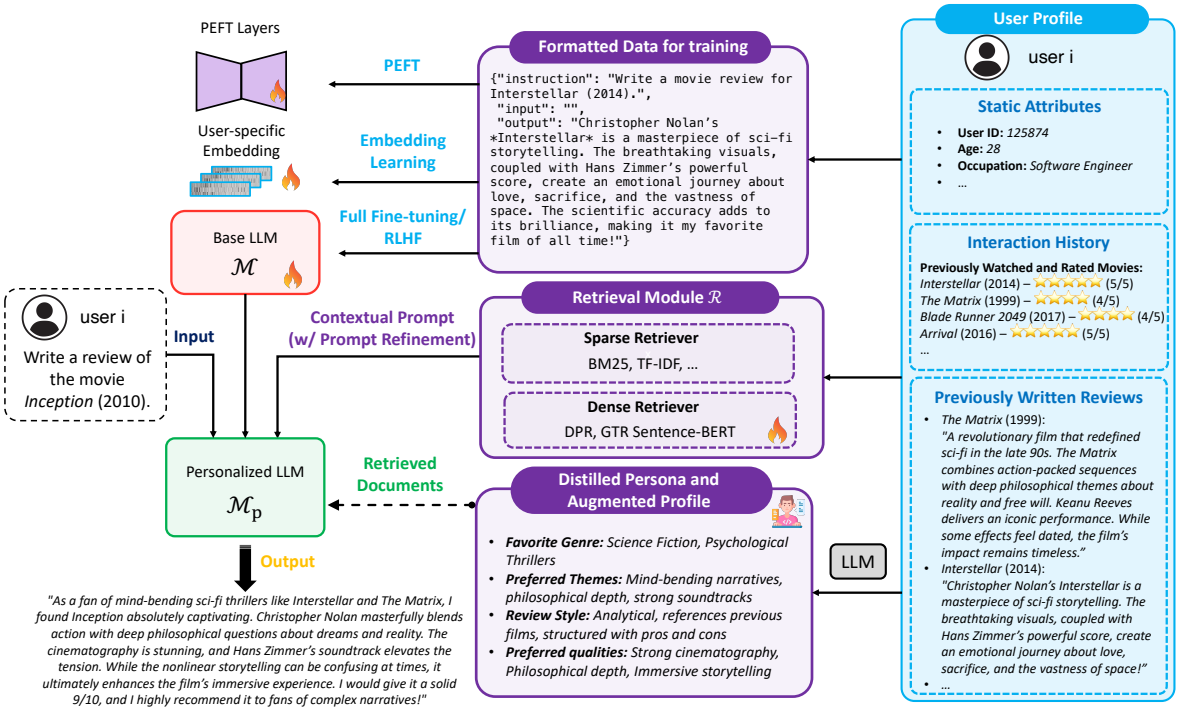


Figure 7: A Case Study on Personalized Movie Review Generation. This figure illustrates how different personalization techniques for LLMs enhance personalized movie review generation by leveraging user profiles, retrieval modules, and fine-tuning methods to align outputs with individual preferences and writing styles.

$$\hat{y}_i = M(\bar{x})$$

where ϕ_q is a query construction function used by the retrieval model $\mathcal{R}$ to find relevant documents, and ϕ_p is a prompt construction function that integrates the retrieved information into the original input X_u. The output $\hat{y}_u$ represents the generated text based on both the original input and the retrieved information.

For personalization tasks, large user profiles often serve as external knowledge bases since they cannot be fully incorporated into prompts due to LLMs' context limitations. As a result, RAG is commonly employed in personalized LLM systems. In this section, we discuss and categorize the personalization techniques that utilize RAG. We categorize those RAG-based personalization techniques based on the retriever into the following two main categories:

- **Sparse Retrieval (Sec. 5.1.1)**: This category of methods employs frequency-based vectors to encode queries and user information, which are then used for retrieval in downstream personalization tasks. Since this approach only requires statistical computations such as frequency counts, it is highly efficient. These methods demonstrate robust performance in information retrieval tasks, frequently serving as baselines in RAG systems.
- **Dense Retrieval (Sec. 5.1.2)**: This category of methods employs deep neural networks including LLM-based encoders to generate continuous embeddings for queries and documents in retrieval tasks. These encoding layers can either use off-the-shelf models directly for downstream tasks without tuning their parameters or incorporate trainable parameters that can be adjusted specifically for retrieval tasks.

Another retrieval method, such as black-box retrieval, involves using external APIs like Google or Bing, which are commonly integrated into LLM-based agent frameworks via tool-use. While this can be valuable for personalization in specific scenarios, we do not explore it in detail due to its black-box nature, which limits transparency regarding how user information is utilized and how personalization is achieved. Additionally,

this design tends to be highly tool-specific, reducing its generalizability. It is also worth noting that many approaches also employ hybrid methods that combine elements of both sparse and dense retrieval.

5.1.1 Sparse Retrieval

Sparse retrieval encodes both queries and documents as sparse vectors, typically based on word frequency and importance. It operates by matching terms in the query with terms in the document, focusing on exact term overlap. Due to its simplicity and effectiveness, sparse retrieval has long been a foundational approach in information retrieval systems. The two most commonly used examples of sparse retrievers are TF-IDF (term frequency-inverse document frequency) (Sparck Jones, 1972) and BM25 (Best Matching 25) (Robertson et al., 1995).

TF-IDF: This method scores documents based on the frequency of terms relative to their occurrence across the entire document collection. It is calculated as:

$$\text{TF-IDF}(q_i, D) = \text{TF}(q_i, D) \cdot \text{IDF}(q_i)$$

where the term frequency (TF) of term q_i in document D is:

$$\text{TF}(q_i, D) = \frac{f(q_i, D)}{|D|}$$

and the inverse document frequency (IDF) is:

$$\text{IDF}(q_i) = \log \frac{N}{n(q_i)}$$

Here, $f(q_i, D)$ is the frequency of term q_i in document D, $|D|$ is the total number of terms in D, N is the total number of documents in the collection, and $n(q_i)$ is the number of documents containing the term q_i. To prevent division by zero and dampen the effect of very rare terms, different smoothed versions are often used.

BM25: This is a more advanced sparse retrieval method that extends the TF-IDF model by incorporating document length normalization and saturation controls for term frequency. The BM25 score for a document D with respect to a query q_i is calculated as:

$$\text{BM25}(q_i, D) = \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{\text{avgdl}})}$$

where q_i is the i-th query term, $f(q_i, D)$ is the term frequency of q_i in document D, $|D|$ is the length of document D, avgdl is the average document length in the collection, and k_1 and b are parameters that control term frequency saturation and document length normalization, respectively. Typical values are $k_1 = 1.2$ and $b = 0.75$.

Because of their generalizability, effectiveness, and simplicity, sparse retrievers often serve as baselines in retrieval-based personalization methods (Salemi et al., 2023; Li et al., 2023b; Richardson et al., 2023). For instance, Salemi et al. (2023) use BM25 as one of the retrievers to fetch relevant user information, which is then incorporated into prompts for LLMs when evaluating on the LaMP dataset (Salemi et al., 2023). Richardson et al. (2023) enhance personalization in LLMs by integrating BM25 retrieval with user data summaries, achieving improved performance on LaMP tasks while reducing the volume of retrieved data by 75%. In another work, Li et al. (2023b) propose a multistage, multitask framework inspired by writing education to enhance personalized text generation in LLMs. In the initial retrieval stage, BM25 is used to retrieve relevant past user documents, which are subsequently ranked and summarized to generate personalized text.

Sparse retrieval serves as a foundation for many personalized systems, particularly in scenarios where large amounts of user information are involved, and efficiency is crucial. However, while sparse retrieval techniques

like BM25 and TF-IDF excel in general retrieval tasks, they have inherent limitations in personalization. The lexical matching nature of these methods struggles to capture semantic relationships between terms, which can hinder their performance in complex personalization tasks. This issue is especially relevant in scenarios where user preferences or behaviors require deeper understanding beyond keyword overlap.

5.1.2 Dense Retrieval

Dense retrievers leverage deep neural networks to generate continuous representations for queries and documents, enabling retrieval in a dense embedding space via similarity-based search (Johnson et al., 2021). Some works (Sun et al., 2024) employ pre-trained LLM encoders, such as OpenAI's `text-embedding-ada` series and Sentence-BERT (Reimers & Gurevych, 2019), without fine-tuning their parameters. Other approaches focus on training retrieval-oriented embeddings. For example, Dense Passage Retrieval (DPR) (Karpukhin et al., 2020) employs dense embeddings in a dual-encoder framework, using BM25 hard negatives and in-batch negatives to efficiently retrieve relevant passages for open-domain question answering. Contriever (Izacard et al., 2021) is an unsupervised dense retriever trained using contrastive learning, where two random spans from a document are cropped independently to form positive pairs for training. In the context of personalization, several works have proposed specialized training data construction (Mysore et al., 2023a;b) and training strategies (Salemi et al., 2023; 2024) to enhance dense retrievers' ability to retrieve more relevant user information, improving personalization for downstream tasks using LLMs. Salemi et al. (2023) use Fusion-in-Decoder (Izacard & Grave, 2021) on encoder-decoder models such as T5 (Raffel et al., 2020) which retrieves and concatenate multiple relevant documents' encoded embedding before decoding. Mysore et al. (2023a) train a pre-trained MPNET model (Song et al., 2020) with a scale-calibrating KL-divergence objective which shows superior performance on personalized open-ended long-form text generation tasks. Salemi et al. (2024) train the Contriever model for personalizing LLMs using LLMs' feedback with policy gradient optimization and knowledge distillation. Zeng et al. (2023) introduce UIA, a flexible dense retrieval framework that incorporates personalized attentive networks to enhance various information access such as keyword search, query by example, and complementary item recommendation. Other dense retrievers such as Sentence-T5 (Ni et al., 2021) and Generalizable T5-based dense Retrievers (GTR) (Ni et al., 2022) are also frequently used for downstream personalization tasks. Generally, while dense retrievers require training on downstream tasks, making them more costly and time-inefficient (Richardson et al., 2023), they tend to achieve superior performance compared to sparse retrievers in downstream personalization tasks. However, constructing effective training data, designing suitable loss functions, and incorporating LLMs into the training process to optimize retrievers for improved downstream personalization tasks remain open challenges.

5.2 Personalization via Prompting

Definition 14 (Prompt Engineering)**.** Prompt Engineering *is the process of designing, refining, and optimizing prompts to achieve desired outputs from language models. This involves iterative testing and adjustment of prompts to enhance the model's performance on various tasks, improve response accuracy, and align the model's outputs with user expectations or specific application requirements.*

A prompt serves as an input for a Generative AI model, guiding the content it generates (Meskó, 2023; White et al., 2023; Heston & Khun, 2023; Hadi et al., 2023; Brown et al., 2020; Schulhoff et al., 2024). Empirically, better prompts enhance LLMs' performances across a wide range of tasks (Wei et al., 2022b; Liu et al., 2023c). As a result, there has been a substantial increase in research dedicated to designing more effective prompts to achieve better outcomes, a field known as prompt engineering. In this section, we categorize personalization techniques that leverage prompt engineering into three main categories:

- **Contextual Prompting (Sec. 5.2.1)**: These methods directly incorporate user history information into the prompt, enabling LLMs to perform downstream personalization tasks based on this contextual data.
- **Persona-based Prompting (Sec. 5.2.2)**: These approaches introduce specific personas, such as demographic information, into the prompt. By encouraging LLMs to role-play these personas, it aims to enhance the performance of downstream personalization tasks.

- **Profile-Augmented Prompting (Sec. 5.2.3)**: These methods focus on designing prompting strategies that enrich the original user history information by leveraging LLMs' internal knowledge, thereby improving downstream personalization tasks.
- **Prompt Refinement (Sec. 5.2.4)**: This category of methods focuses on developing robust frameworks that iteratively refine the initial hand-crafted prompts, enhancing downstream personalization.

5.2.1 Contextual Prompting

As current LLMs demonstrate increasing abilities and extended context lengths (Jin et al., 2024a; Ding et al., 2024; Lin et al., 2024b), one naive approach is to directly include a proportion of past user information in the prompt and ask the LLMs to predict user behavior on downstream tasks (Di Palma et al., 2023; Wang & Lim, 2023; Sanner et al., 2023; Li et al., 2023e; Christakopoulou et al., 2023). For example, Kang et al. (2023) investigate the performance of multiple LLMs in user rating prediction tasks by directly incorporating the user's past rating history and candidate item features in a zero-shot and few-shot manner. This work finds that LLMs underperform traditional recommender systems in zero-shot settings but achieve comparable or superior results when fine-tuned with minimal user interaction data. Larger models (100B+ parameters) show better performance and faster convergence, highlighting LLMs' data efficiency and potential in recommendation tasks. Similarly, Liu et al. (2023a) investigate the potential of ChatGPT as a general-purpose recommendation model by directly injecting user information in the prompt and evaluating its performance on five recommendation tasks: rating prediction, sequential recommendation, direct recommendation, explanation generation, and review summarization. The study finds that while ChatGPT performs well in generating explanations and summaries, it shows mixed results in rating prediction and poor performance in sequential and direct recommendations, indicating the need for further exploration and improvements. These studies suggest that directly incorporating past user information into LLM prompts as contextual input could be a promising solution for a wide range of personalized downstream tasks. However, this approach faces challenges in scalability when dealing with large, unstructured user data, as LLMs may struggle to interpret such data effectively (Liu et al., 2024b). Additionally, while it offers improved explainability, it may not achieve significant performance gains over traditional non-LLM-based methods.

5.2.2 Persona-based Prompting

LLMs have been widely used for role-playing and imitating human behavior, mainly by specifying the desired persona within the prompt (Aher et al., 2023; Horton, 2023; Kovač et al., 2023; Argyle et al., 2023; Dillion et al., 2023; Woźniak et al., 2024; Li et al., 2024e). Generally, the persona denotes the entity whose viewpoints and behaviors the simulation seeks to examine and reproduce. This persona can encompass relatively stable characteristics (e.g., race/ethnicity), those that gradually evolve (e.g., age), or those that are transient and situational (e.g., emotional state) (Yang, 2019). Chen et al. (2024b) categorize persona into three types: *demographic persona*, which represents aggregated characteristics of demographic segments (Huang et al., 2023a; Xu et al., 2023; Gupta et al., 2023); *character persona*, encompassing well-established characters from real and fictional sources (Shao et al., 2023; Wang et al., 2023d; 2024f); and *individualized persona*, constructed from individual behavioral and preference data to provide personalized services which is discussed in Sec. 5.2.1. For example, to induce an extroverted persona in LLMs, Jiang et al. (2023) use the prompt of" *You are a very friendly and outgoing person who loves to be around others. You are always up for a good time and love to be the life of the part*" which achieve more consistent results on human psychological assessments such as the Big Five Personality Test (Barrick & Mount, 1991). Through this type of prompt construction, LLMs can deviate from their intrinsic "personality" (Karra et al., 2022; Safdari et al., 2023; Huang et al., 2023b; Santurkar et al., 2023; Hartmann et al., 2023) and exhibit altered characteristics in their responses in accordance with the requirements in the prompt. Another approach involves using LLMs to mimic well-known figures (e.g., Elon Musk). Prior works achieve this by incorporating descriptions of character attributes—such as identity, relationships, and personality traits—or by providing demonstrations of representative behaviors that reflect the characters' linguistic, cognitive, and behavioral patterns within the prompt (Han et al., 2022; Li et al., 2023a; Chen et al., 2023a; Zhou et al., 2023; Shen et al., 2023; Yuan et al., 2024; Chen et al., 2024a). While persona-based role-playing can effectively reflect certain personalities, potentially improving adaptive personalization by dynamically adjusting to user-specific behaviors and preferences over time, it also raises

significant concerns. 'Persona prompting' may lead to issues such as 'character hallucination,' where the model exhibits knowledge or behavior misaligned with the simulated persona. Additionally, it may introduce biases (Gupta et al., 2023; Zhang et al., 2023d; Wang et al., 2024a; Ziems et al., 2024), toxicity (Deshpande et al., 2023), potential jailbreaking (Chao et al., 2023; Liu et al., 2023g; Chang et al., 2024; Xu et al., 2024b), ecological fallacy (Orlikowski et al., 2023), and susceptibility to caricatures (Cheng et al., 2023b), among other risks.

5.2.3 Profile-Augmented Prompting

In many personalization datasets, there are two main issues with the user-profile database. First, the size of the user data is often so large that it may exceed the model's context length or contain a significant amount of irrelevant information, which can distract the model (Shi et al., 2023; Liu et al., 2024c). Second, despite the large size, the user-profile database frequently contains incomplete or insufficient information (Perez et al., 2007; Dumitru et al., 2011) and sparse user interactions (e.g., cold-start). For example, movie recommendation datasets typically only contain the main actors and a brief plot summary, but this often overlooks key details like genre, tone, and thematic depth, leading to less effective recommendations. Another example could be 'lurkers,' users with minimal interaction history, a common scenario in recommendation systems that makes it difficult to give personalized responses. To resolve these problems, a line of work focuses on eliciting the internal knowledge of LLMs to augment or distill existing user profiles (Zheng et al., 2023b; Wu et al., 2024b). Richardson et al. (2023) propose a summary-augmented approach by extending retrieval-augmented personalization with task-aware user summaries generated by LLMs in the prompt. ONCE (Liu et al., 2024d) generates user profiles by prompting LLMs to summarize the topics and regions of interest extracted from their browsing history, which helps LLMs capture user preferences for downstream tasks. Lyu et al. (2023) propose LLM-REC, which employs four prompting strategies to augment the original item descriptions, which often contain incomplete information for recommendation. These augmented descriptions are then concatenated as the input for the following recommendation module, introducing the relevant context and helping better align with user preferences. Xi et al. (2023) use factorization prompting to extract nuanced user preferences and item details from user profiles, and employs a hybrid-expert adaptor to convert this knowledge into augmented vectors for existing recommendation models. Sun et al. (2024) introduce Persona-DB, a hierarchical construction of user personas from interaction histories through prompting LLMs, and a collaborative refinement process that integrates data from similar users to enhance the accuracy and efficiency of personalized response forecasting.

5.2.4 Prompt Refinement

Most works using prompt engineering in personalization tasks rely on hand-crafted prompts, which necessitate human expertise and can be costly, with their effectiveness being verifiable only through trial and error. Some research efforts aim to train models to refine these manually designed prompts, enhancing their capability for personalization. In the context of personalization, Li et al. (2024a) train a small LM such as T5 (Raffel et al., 2020) for revising text prompts and enhancing personalized text generation with black-box LLMs. Their approach combines supervised learning and reinforcement learning to optimize prompt rewriting. Kim & Yang (2024) propose FERMI (Few-shot Personalization of LLMs with Mis-aligned Responses), a method that optimizes input prompts by iteratively refining them based on user profiles and past feedback, while also incorporating contexts of misaligned responses to enhance personalization. The optimization process involves three steps: scoring the prompts based on user feedback, updating a memory bank of high-scoring prompts and their contexts, and generating new, improved prompts. Additionally, the method refines inference by selecting the most relevant personalized prompt for each test query, leading to significant performance improvements across various benchmarks (Santurkar et al., 2023; Durmus et al., 2023; Salemi et al., 2023).

5.3 Personalization via Representation Learning

Personalized representation learning aims to learn latent representations that accurately capture each user's behavior, with applications in personalized response generation, recommendations, etc (Li & Zhao, 2021; He et al., 2023; Tan & Jiang, 2023). In this section, we discuss and categorize personalization techniques that leverage representation learning into the following main categories:

- **Full-Parameter Fine-tuning (Sec. 5.3.1)**: This category of methods focuses on developing training strategies and curating datasets to update all parameters of the LLM, enhancing its ability to perform downstream personalization tasks more effectively.
- **Parameter-Efficient Fine-tuning (PEFT) (Sec. 5.3.2)**: This category of methods avoids fine-tuning all the parameters by updating only a small number of additional parameters or a subset of the pretrained parameters to adapt the LLMs to downstream personalization tasks. This selective tuning allows for the efficient encoding of user-specific information.
- **Embedding Learning (Sec. 5.3.3)**: This category of methods focuses on learning embeddings that represent both input text and user information in vectorized form, enabling models to more effectively incorporate personalized features and preferences into the learning process.

5.3.1 Full-Parameter Fine-tuning

Definition 15 (Fine-tuning)**.** Fine-tuning *is the process of adapting a LLM $\mathcal{M}$ to a specific task by further training it on a smaller, targeted dataset $\mathcal{D}_i$ after the pre-training stage. This updates the model's parameters θ to improve performance on the specified downstream tasks. Formally, fine-tuning can be expressed as:*

$$\theta^* = \arg\min_{\theta} \mathcal{L}(\mathcal{M}(X_i; \theta), Y_i)$$

where X_i is the input data, and Y_i is the corresponding output. The fine-tuned model $\mathcal{M}$ with parameters θ^* is optimized for generating desired responses.

The "pre-train, then fine-tune" paradigm has been widely adopted, enabling the development of foundation models that can be adapted to a range of applications after acquiring general knowledge from large corpus pre-training (Bommasani et al., 2021; Min et al., 2023; Liu et al., 2023c). Fine-tuning LLMs for targeted scenarios generally yields better results on most tasks when model parameters are accessible and the associated costs are acceptable (Gao et al., 2023). For instance, fine-tuning allows LLMs to adapt to specific data formats and generate responses in a particular style as instructed, which is crucial for many personalization tasks (Du & Ji, 2022). Empirically, it can achieve better performances on some personalization tasks than zero-shot or few-shot prompting off-the-shell LLMs (Kang et al., 2023). Li et al. (2023b) use an auxiliary task, called author distinction, to train a T5-11B model (Raffel et al., 2020) to better distinguish if two documents are from the same user to obtain a better-personalized representation. Yin et al. (2023) first use ChaGPT to fuse heterogeneous user information through prompt engineering to build an instructional tuning dataset and use this dataset to fine-tune ChatGLM-6B (Du et al., 2021), resulting in enhanced recommendation performance. In this line of work, a common approach involves providing a task instruction that includes user interaction history and a potential item candidate, with a label of 'Yes' or 'No' indicating whether the user prefers the item. Fine-tuning LLMs, such as LLaMA (Touvron et al., 2023), on such instruction-tuning datasets generally yields superior performance on recommendation tasks compared to both prompting LLMs and traditional methods (Hidasi et al., 2015). Yang et al. (2023) fine-tune a LLaMa 7B with an instruction tuning dataset that provides instructions for generating a list of "future" items the user may interact with, based on a list of past interactions, or for retrieving target "future" items from a list of candidate items.

5.3.2 Parameter-Efficient Fine-tuning

Definition 16 (Parameter-efficient Fine-tuning)**.** Parameter-efficient Fine-tuning *(PEFT) is a technique for adapting LLMs to specific tasks by updating only a small subset of the model's parameters $\theta_t \subset \theta$ or introducing a limited set of new parameters θ_{new}, while keeping the majority of the original parameters θ unchanged.*

Formally, given an LLM $\mathcal{M}$ with parameters θ, PEFT seeks to minimize the loss function $\mathcal{L}$ over a reduced set of parameters:

$$\theta_t^*, \theta_{new}^* = \arg\min_{\theta_t, \theta_{new}} \mathcal{L}\left(\mathcal{M}(X_i; \theta_t, \theta_{new}, \theta_{frozen}), Y_i\right)$$

where X_i is the input data, Y_i is the corresponding output, and θ_{frozen} represents the parameters that remain fixed during fine-tuning, and θ_{new} may be empty if no new parameters are introduced.

Tan et al. (2024b) introduce One PEFT Per User (OPPU), a method that employs personalized PEFT modules such as LoRA (Hu et al., 2021) and prompt tunning parameters (Lester et al., 2021) to encapsulate user-specific behavior patterns and preferences. Based on OPPU, Tan et al. (2024a) further propose PER-PCS, a framework enabling efficient and fine-grained personalization of LLMs by allowing users to share and assemble personalized PEFT pieces collaboratively. Dan et al. (2024) introduce P-Tailor which personalizes LLMs by using a mixture of specialized LoRA experts to model the Big Five Personality Traits. Huang et al. (2024b) propose Selective Prompt Tuning which improves personalized dialogue generation by adaptively selecting suitable soft prompts for LLMs based on input context.

5.3.3 Embedding Learning

Definition 17 (Embedding)**.** *An* Embedding *is a vector in a continuous space produced by an embedding function $Emb(\cdot)$, which maps discrete data, such as tokens, into continuous vector spaces. This transformation allows text to be represented in a format suitable for machine learning models. Given a token w, the embedding function produces a vector $\mathbf{e} \in \mathbb{R}^d$, where d is the dimensionality of the embedding space:*

$$\mathbf{e} = Emb(w)$$

Definition 18 (User-specific Embedding Learning)**.** User-specific Embedding Learning *involves creating embeddings that capture individual user preferences and behaviors from their interaction data. These embeddings are used to personalize model outputs.*

Formally, given user interactions I_i, the embedding $\mathbf{e}_i$ is obtained via:

$$\mathbf{e}_i = Emb(I_i)$$

The embedding is then integrated into the model to generate personalized responses:

$$g_i = \mathcal{M}(X_i; \mathbf{e}_i)$$

where X_i is the input data, and $\mathcal{M}$ is the model. This approach enhances personalization by adapting the model's responses to individual user characteristics.

Cho et al. (2022) present a personalized dialogue generator that detects an implicit user persona using conditional variational inference to produce user-specific responses based on dialogue history, enhancing user engagement and response relevance. HYDRA (Zhuang et al., 2024) enhances black-box LLM personalization by capturing user-specific behavior patterns and shared general knowledge through model factorization. It employs a two-stage retrieve-then-rerank workflow and trains an adapter to align model outputs with user-specific preferences. Ning et al. (2024b) propose a USER-LLM that leverages user embeddings to dynamically contextualize LLMs. These user embeddings are distilled from diverse user interactions using self-supervised pretraining, capturing latent user preferences and their evolution over time. By integrating these embeddings through cross-attention and soft-prompting, the approach enhances personalization and performance across various tasks while maintaining computational efficiency. Liu et al. (2024a) propose the PPlug model, which personalizes LLMs by employing a lightweight plug-in user embedder that aggregates user historical behaviors into a single, input-aware embedding. This embedding guides a fixed LLM to generate personalized outputs without modifying the model's parameters, significantly improving personalization performance over retrieval-based methods.

5.4 Personalization via Reinforcement Learning from Human Feedback (RLHF)

LLMs are learned in multiple stages (Ouyang et al., 2022): pre-training on large amounts of text, fine-tuning on domain-specific data, and alignment to human preferences. While the alignment is usually done to capture general user preferences and needs, it can also be used for personalization, to align to expectations and requirements of individual users. The set of algorithmic techniques used for alignment is know as *Reinforcement Learning from Human Feedback (RLHF)* and we review these techniques next.

In classic *reinforcement learning (RL)* (Sutton & Barto, 1998), an agent learns a policy to optimize a long-term goal from reward signals. This can be done by directly learning a policy (Williams, 1992; Baxter & Bartlett,

2001; Schulman et al., 2015; 2017), learning a value function (Bellman, 1957; Sutton, 1988), or a combination of both (Sutton et al., 2000). In RLHF, the agent learns a policy represented by an LLM (Ouyang et al., 2022; Ahmadian et al., 2024; Rafailov et al., 2024; Xu et al., 2024a), which is optimized based on preferential human feedback on its outputs. The preferential feedback is rooted in social sciences, and can be binary (Bradley & Terry, 1952) or over multiple options (Plackett, 1975; Zhu et al., 2023a). RLHF helps the LLM to align with human values, promoting more ethically sound and socially responsible AI systems (Kaufmann et al., 2023). The first methods for aligning LLMs through RLHF were based on learning a proxy reward model from human feedback (Nakano et al., 2021; Ouyang et al., 2022; Bai et al., 2022; Dubois et al., 2024; Lin et al., 2024a; Chakraborty et al., 2024). This can be viewed as a form of reward shaping (Ng et al., 1999) applied to a reward model that captures general preferences of a population. Modern alignment techniques (Rafailov et al., 2024) optimize the LLM directly from human feedback.

Besides general preferences, some works (Jang et al., 2023) also investigate the alignment of LLMs with personalized human preferences. The motivation for this task is that even for the same prompt, different users may desire different outputs, and individual preferences can vary across different dimensions (Casper et al., 2023). For example, when asked "What is an LLM?", a PhD student in NLP may prefer a detailed technical explanation, while a non-expert might seek a simplified and concise definition. Jang et al. (2023) frame this problem as a Multi-Objective Reinforcement Learning (MORL) task, where diverse and potentially conflicting user preferences are decomposed into multiple dimensions and optimized independently. It can be efficiently trained independently and combined effectively post-hoc through parameter merging. Chen et al. (2024d) introduce Personalized Alignment at Decoding-Time (PAD), a training-free framework for aligning large language models with personalized user preferences during inference. PAD employs a personalized reward modeling strategy that decouples text generation dynamics from user preferences, enabling generalizable token-level personalized rewards to guide the decoding process. This approach dynamically adjusts model outputs to diverse or unseen preferences without requiring retraining, demonstrating scalability and superior alignment performance across multiple base models. Li et al. (2024g) propose a Personalized-RLHF (P-RLHF) framework, where user-specific models are jointly learned alongside language or reward models to generate personalized responses based on individual user preferences. The method includes developing personalized reward modeling (P-RM) and personalized Direct Preference Optimization (P-DPO) objectives, tested on text summarization data, demonstrating improved alignment with individual user-specific preferences compared to non-personalized models. Park et al. (2024a) address the challenge of heterogeneous human preferences in RLHF, using personalized reward models through representation learning and clustering, as well as preference aggregation techniques grounded in social choice theory and probabilistic opinion pooling. Kirk et al. (2024) introduce the PRISM Alignment Project which is a novel dataset that maps the sociodemographics and preferences of 1,500 diverse participants from 75 countries onto their feedback in over 8,000 live conversations with 21 LLMs. This dataset aims to enhance the alignment of AI systems by incorporating wide-ranging human perspectives, especially on value-laden and controversial topics. Lee et al. (2024) propose an approach to aligning LLMs with diverse human preferences through system message generalization, leveraging a comprehensive dataset named Multifaceted Collection to train the model JANUS, which effectively adapts to a wide range of personalized user preferences. Yang et al. (2024a) propose Rewards-in-Context which fine-tunes LLMs by conditioning responses on multiple rewards during supervised fine-tuning, enabling flexible preference adaptation at inference time. This method achieves Pareto-optimal alignment across diverse objectives, using significantly fewer computational resources compared to traditional MORL approaches. Poddar et al. (2024) propose Variational Preference Learning (VPL), which is a technique to personalize RLHF by aligning AI systems with diverse user preferences through a user-specific latent variable model. VPL incorporates a variational encoder to infer a latent distribution over hidden user preferences, enabling the model to condition its reward functions and adapt policies based on user-specific context. In simulated control tasks, VPL demonstrates effective modeling and adaptation to diverse preferences, with enhanced performance and personalization capabilities compared to standard RLHF methods. Another idea is to offer demonstrations as an efficient alternative to pairwise preferences, allowing users to directly showcase their desired behavior through example completions or edits. This method, as demonstrated by DITTO (Shaikh et al., 2024), enables rapid and fine-grained alignment to individual preferences with minimal data, making it a potentially powerful tool for personalizing LLMs with limited user input.

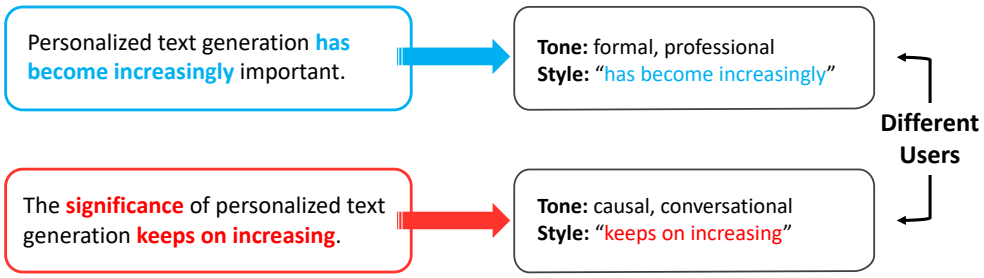


Figure 8: **Personalization of User-specific Writing Style and Tone.**

5.5 Discussion

Here, we discuss the computational and scalability considerations, effectiveness comparisons, and practical guidance for implementing various personalization techniques for LLMs. From a computational and scalability perspective, personalization strategies vary significantly in their resource demands and implementation complexity. RAG-based methods, especially when relying on dense retrievers, can deliver strong personalization fidelity (Salemi et al., 2023) but may incur considerable indexing overhead and GPU memory usage (Asai et al., 2023; Fan et al., 2024). This has led to growing interest in hybrid approaches that integrate sparse methods (e.g., BM25) for greater efficiency with dense retrieval techniques to maintain semantic alignment (Novotnỳ & Stefánik, 2022; Arabzadeh et al., 2021). Prompting-based methods generally do not require large-scale databases or model parameter tuning, making them computationally efficient and lightweight. However, deriving effective personas from extensive user data or implementing multi-level hierarchical prompting (Sun et al., 2024) may involve iterative model inference, introducing latency challenges in real-world applications. In the future, distillation-based approaches (Li et al., 2023c), which transfer the general capabilities of large LLMs into smaller, task-specific models tailored for personalization, could potentially enhance efficiency by preserving personalization fidelity while reducing computational costs. Other techniques such as caching pipelines (Wu et al., 2024c) and on-demand personalization for high-value users also shows promising as practical solutions to reduce inference latency and memory consumption. Additionally, federated or on-device personalization offers a promising avenue to alleviate server-side resource burdens while ensuring user privacy (Kulkarni et al., 2020; Li et al., 2021). RLHF-based methods, while capable of delivering high-quality personalization, often require extensive and costly human feedback collection, which can be prone to noise (Casper et al., 2023; Kaufmann et al., 2023). Alternatives such as Reinforcement Learning from AI Feedback (RLAIF) (Lee et al.) may reduce costs but introduce challenges, such as limitations in capturing nuanced user preferences and ensuring feedback quality. In terms of effectiveness, recent benchmarks like LamP (Salemi et al., 2023) and longLamP (Kumar et al., 2024) indicate that methods leveraging user-specific data retrieval (e.g., BM25, Contriever) generally excel in tasks that demand explicit grounding, with dense retrievers proving especially adept at complex abstractions. Prompt-based personalization remains highly flexible and lightweight in scenarios with moderate amounts of user data, yet its performance degrades as context windows become saturated. Meanwhile, Kumar et al. (2024) find that representation learning approaches such as fine-tuning FLAN-T5 or incorporating user embeddings tend to yield deeper personalization gains, though they risk overfitting when per-user data is sparse and require substantial computational resources for training. As a result, deciding which technique to employ depends largely on application requirements and constraints. RAG is well-suited to large knowledge bases or dynamic user profiles where context must be retrieved on the fly, while prompting methods are preferable for conversational systems that can accommodate brief historical interactions without retraining. Organizations with abundant, high-quality user data and significant computational capacity may consider representation learning, particularly when nuanced personalization is important. RLHF demonstrates significant promise in scenarios where continuous alignment with user feedback is essential, particularly when pairwise preference data is readily available. However, the choice of personalization method is shaped by multiple factors, including the availability of data, infrastructure constraints, latency requirements, and regulatory considerations. These diverse influences ensure that no single personalization approach is universally optimal across all tasks, highlighting the need for context-specific strategies.

6 Taxonomy of Evaluation Metrics for Personalized LLMs

In this section, we introduce a taxonomy for the evaluation of personalized LLM techniques. In particular, we categorize evaluation as *intrinsic evaluation* of the personalized text generated (Sec. 6.1) or *extrinsic evaluation* that relies on a downstream application such as recommendation systems to demonstrate the utility of the generated text from the personalized LLM (Sec. 6.2).

The performance of a *personalized LLM* for a downstream application (such as the generation of an email personalized for a specific user or the generation of an abstract by a specific user) should be quantified using an appropriate evaluation metric. As an example, for the direct evaluation task of personalized text generalization, there are many facets that must be considered for the personalized evaluation of the generated text for a specific user. There are largely the following factors that are illustrated in Figure 4 and Table 1. Figure 8 illustrates an example on how the same information can be personalized and expressed in different writing styles and tones, which should be measured by the employed metrics.

Definition 19 (Evaluation Metric)**.** *For an arbitrary dataset $\mathcal{D}$, there is a subset of* evaluation metrics *$\psi(\mathcal{D}) \subseteq \Psi$ that can be used for $\mathcal{D}$, where Ψ is the space of all metrics and $\psi(\mathcal{D})$ is the subset of metrics appropriate for the dataset $\mathcal{D}$.*

Definition 20 (Intrinsic Evaluation)**.** Intrinsic Evaluation $\mathbb{E}_i$ *refers to the assessment of personalized text generated by an LLM $\mathcal{M}_p$ based on predefined metrics $\psi(\cdot) \in \Psi$ that measure the quality, relevance, and accuracy of the generated content $\hat{y} \in \hat{\mathbb{Y}}$ against ground-truth data $Y \in \mathbb{Y}$. This evaluation is performed directly on the output of the model:*

$$\mathbb{E}_i(\hat{y}, Y) = \psi(\hat{y}, Y)$$

Note that the ground-truth data Y can represent user-written text when available or, alternatively, user preferences and reward models that capture user judgments.

Definition 21 (Extrinsic Evaluation)**.** Indirect Evaluation $\mathbb{E}_e$ *involves assessing the utility of the personalized text generated by an LLM $\mathcal{M}_p$ through its impact on a downstream application $\mathcal{F}$. The evaluation measures the effectiveness of the generated content by comparing the predictions $\hat{r}$ with the ground-truth labels r using application-specific metrics:*

$$\mathbb{E}_e(\hat{r}, r) = \psi_a(\hat{r}, r)$$

where $\psi_a(\cdot) \in \Psi_a$ represents the application-specific metrics.

More formally, let $\mathbb{E}_i$ represent intrinsic evaluation and $\mathbb{E}_e$ represent extrinsic evaluation. Let $\hat{y} = \mathcal{M}(X; \theta)$ denote the generated content for input X from dataset D, and let Y represent the ground truth output from D. Similarly, let $\hat{r} = \mathcal{F}(\hat{y})$ be the downstream task predictions based on $\hat{y}$, and r be the ground truth output for the downstream tasks.

- **Intrinsic Evaluation (Sec. 6.1)**: Formally, intrinsic evaluation metrics can be defined as $\psi(\mathcal{D}) = \{\mathbb{E}_i \mid \mathbb{E}_i(\hat{y}, Y)\}$. Most research on personalized text generation focuses on scenarios where ground-truth user-written text is available (Salemi et al., 2023; Kumar et al., 2024). In this setting, common evaluation metrics include BLEU (Papineni et al., 2002), ROUGE-1 (Lin & Hovy, 2003), ROUGE-L (Lin & Och, 2004), METEOR (Banerjee & Lavie, 2005), BERTScore (Zhang et al., 2019), and Hits@K. BLEU is primarily used for text generation tasks, such as machine translation. ROUGE-1 and ROUGE-L belong to the ROUGE family (Lin, 2004), originally designed for summarization evaluation. ROUGE-1 measures unigram recall between the predicted and reference summaries, while ROUGE-L considers the longest common subsequence between them. METEOR, originally developed for machine translation evaluation, focuses on string alignment. BERTScore measures the similarity between contextual embeddings generated by the BERT model (Devlin et al., 2019). Hits@k measures the percentage of test cases where the correct answer appears in the top k predictions. For example, with the persona "I love hiking" and the correct response "I hike every weekend" ranked first, it counts as a hit for Hits@1, Hits@3, and Hits@5. Higher scores in these metrics indicate better model performance. However, there are scenarios where ground-truth text is not available. In such cases, model alignment with human pairwise preferences or reward models that capture user intent can serve as alternatives. However, this aspect has not been systematically explored yet.

- **Extrinsic Evaluation (Sec. 6.2)**: Extrinsic evaluation metrics can be expressed as $\psi(\mathcal{D}) = \{\mathbb{E}_e \mid \mathbb{E}_e(\hat{r}, r)\}$. These metrics are used to evaluate the generated content based on its effectiveness in downstream tasks, such as recommendation or classification. For recommendation tasks, common metrics include Recall, Precision, and Normalized Discounted Cumulative Gain (NDCG). In typical recommendation systems, the top-k items are returned by personalized LLMs. Recall and Precision evaluate whether the predicted top-k items match the expected top-k items, while NDCG takes into account the ranking of the recommendations. For classification tasks, metrics such as Recall, Precision, Accuracy, and F1 Score are frequently used to measure performance.

Table 4 provides a taxonomy of evaluation metrics for personalized LLMs.

6.1 Intrinsic Evaluation

Intrinsic evaluation metrics are primarily used when ground truth textual data is available to assess the quality of generated content. In LaMP (Salemi et al., 2023), BLEU, ROUGE-1, and ROUGE-L are employed to evaluate models on tasks such as personalized news headline generation, personalized scholarly title generation, personalized email subject generation, and personalized tweet paraphrasing. More recently, the LongLaMP (Kumar et al., 2024) benchmark was proposed to evaluate personalized LLM techniques for longer-form personalized text generation. Similarly, ROUGE-1, ROUGE-L, and METEOR metrics are utilized to assess personalized LLMs on tasks like (1) personalized abstract generation, (2) personalized topic writing, (3) personalized review writing, and (4) personalized email writing. In addition, the win rate metric (Hu et al., 2024) is used to evaluate personalized responses for medical assistance (Zhang et al., 2023b), and Hits@K is applied to assess personalized responses in dialogues (Mazaré et al., 2018). Word Mover's Distance is another metric used to evaluate personalized review generation (Li & Tuzhilin, 2019). EGISES (Vansh et al., 2023) is the first metric explicitly designed to assess a summarization model's responsiveness to user-specific preferences. It does so by measuring the degree of insensitivity of model-generated summaries to variations across reader profiles, using Jensen-Shannon divergence to quantify the deviation between expected user-specific summaries and generated summaries. This approach allows EGISES to capture personalization independently of accuracy, establishing a baseline for evaluating how well a summarization model can adapt to individualized preferences beyond mere accuracy. Building on EGISES, PerSEval (Dasgupta et al., 2024) introduces a refined metric that not only assesses personalization but also integrates accuracy considerations to better reflect real-world performance. PerSEval differentiates itself by introducing the Effective DEGRESS Penalty (EDP), which imposes penalties for drops in summary accuracy and inconsistency across summaries. This design balances alignment with user preferences and accuracy, ensuring that high responsiveness does not mask poor accuracy, a limitation in EGISES.

LLMs are increasingly being used as evaluators to reduce the reliance on human labor. For example, MT-bench and Chatbot Arena (Zheng et al., 2023a) use strong LLMs as judges to evaluate these models on open-ended questions. However, questions remain about how reliably LLMs can serve as evaluators. Chiang & Lee (2023) first defined *LLM evaluation* as the process of feeding LLMs the same instructions and questions given to human evaluators and obtaining answers directly from LLMs. Judge-Bench (Bavaresco et al., 2024) provides an empirical study comparing LLM evaluation scores with human judgments, concluding that LLMs are not yet ready to fully replace human judges in NLP tasks due to high variance in their evaluation performance. EvalGen (Shankar et al., 2024) incorporates human preferences by allowing users to grade LLM-generated prompts and code. This approach iteratively refines evaluation criteria based on user feedback, helping to validate the quality of generated content. In the context of personalization, to the best of our knowledge, there is still a lack of an LLM-as-a-judge framework specifically designed to systematically evaluate personalization applications. Developing such a framework or constructing robust reward models that accurately capture diverse user preference could be a promising direction for future research. Additionally, human evaluation remains indispensable in this area. Methods like human preference judgments and pairwise comparisons are commonly used to assess the alignment between generated content and user-specific requirements. While LLMs can assist in the evaluation or annotation process (Li et al., 2023d), human evaluation remains essential for ensuring that personalized outputs truly meet user expectations, despite its practical cost.

6.2 Extrinsic Evaluation

Extrinsic evaluation metrics assess the quality of personalized LLMs in downstream tasks, such as recommendation and classification. For recommendation tasks, a common example is top-k recommendation, where personalized LLMs predict which top-k items to recommend. The predicted recommendations are then compared to the reference items (ground truth). Commonly used metrics include recall, precision, and NDCG. Recall measures the percentage of relevant items retrieved from the reference set, while Precision indicates the percentage of correctly retrieved items within the recommendations. NDCG evaluates how closely the ranking of recommended items matches the ground truth rankings. For other recommendation tasks, such as rating prediction, metrics like mean squared error (MSE), root mean squared error (RMSE), and mean absolute error (MAE) are frequently used.

In classification tasks, personalized LLMs classify input text (such as profiles or descriptions) into one of several candidate classes, which may involve binary or multiclass classification. For instance, personalized LLMs may match patients with suitable clinical trials, where the input includes patient health records and trial descriptions, and the output indicates whether there is a match (Yuan et al., 2023). In this context, metrics such as Recall, Precision, and F1 Score are used to evaluate the quality of the matches. These metrics are similarly applied in other classification tasks, including entity recognition and relation extraction (Tang et al., 2023). Additionally, metrics like accuracy and micro-F1 Score are often employed to assess the quality of classification outputs. For example, personalized LLMs may classify movies into specific categories based on their profiles (Salemi et al., 2023), with accuracy and micro-F1 used to measure classification performance. Beyond the tasks mentioned, a wide range of personalization tasks may require other task-specific metrics for evaluation. While most of these metrics often do not directly assess personalization, improvements often implicitly reflect enhanced personalization capabilities through better utilization of user information.

Table 4: Taxonomy of Evaluation Metrics for Personalized LLMs

Metric	Input format	Output format	Key Equation	Datasets or Domain
TEXT-GENERATION				
BLEU	N-grams	Scalar	$BP \times exp\left(\sum w_n \log p_n\right)$	Email, Amazon (Li et al., 2024b)
ROUGE-1	Unigrams	Scalar	$\frac{\lvert Y_i \cap \hat{Y}_i \rvert}{\lvert Y_i \rvert}$	LaMP (Salemi et al., 2023)
ROUGE-L	Sequences	F-measure	$\frac{(1+\beta^2)Re \times Pr}{Re+\beta Pr}$	LaMP (Salemi et al., 2023)
METEOR	Alignments	F-measure	$\alpha \times \frac{10Re \times Pr}{Re+9Pr}$	LongLaMP (Kumar et al., 2024)
Bertscore	Tokens	Scalar	(Zhang* et al., 2020)	Writing (Mysore et al., 2023a)
Hits@K	Answers	Scalar	$\frac{t \in \mathcal{K}_{test} \wedge t \leq k}{\lvert \mathcal{K}_{test} \rvert}$	Dialogue (Mazaré et al., 2018)
DOWNSTREAM TASKS				
RECOMMENDATION				
Recall	Articles	Recommendations	$\frac{TP}{TP+FP}$@top-k	Product, News (Ni et al., 2019; Wu et al., 2020)
Precision	Comments	Recommendations	$\frac{TP}{TP+FN}$@top-k	Trip, Yelp (Li et al., 2020; Yelp., 2014)
NDCG	Descriptions	Rankings	$\frac{\text{DCG}}{\text{IDCG}}$@top-$k$	Movie, Recipe (Lyu et al., 2023)
CLASSIFICATION				
Recall	Profiles	Classes	$\frac{TP}{TP+FP}$	Healthcare (Yuan et al., 2023)
Precision	Profiles	Classes	$\frac{TP}{TP+FN}$	Citation (Salemi et al., 2023)
Accuracy	Profiles	Classes	$\frac{TP+TN}{TP+TN+FP+FN}$	Category (Salemi et al., 2023)
Micro-F1 Score	Descriptions	Multi-Classes	$\frac{TP}{TP+\frac{1}{2}(FP+FN)}$	Category (Salemi et al., 2023)
LLMS AS EVALUATORS				
IAA	Prompts	Likert Scores	$\alpha = \frac{p_a - p_e}{1-p_e}$	Questions (Chiang & Lee, 2023)
Correlations	Texts	Ratings	$\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2 \sum(y_i-\bar{y})^2}}$	Translation (Bavaresco et al., 2024)
Pass Rate	Prompts	Code	$\sum_{f \in F} \text{PR}(f) \times f(e)$	Medical (Shankar et al., 2024)

Table 5: **Taxonomy of Datasets for Personalized LLMs.** The datasets are categorized by the downstream task they have been applied to, including: **short-text generation**, **long-text generation**, **recommendation**, **classification**, **dialogue**, and **question answering**. Notably, benchmark datasets for short-text and long-text generation contain user-specific ground-truth text, while others primarily rely on task-specific labels in different formats. For each dataset, we also indicate whether data has been filtered to include only users with a substantial amount of prior interactions (e.g., users who have reviewed at least 100 products, sent at least 100 emails, or rated at least k movies). In addition, we identify whether the dataset contains user-generated text, numerical attributes (such as ratings), or other categorical attributes (e.g., the genre of a movie a user watched). While attributes like movie descriptions are text-based, they are not considered user-generated content for the purpose of personalized LLM techniques. Evaluation metrics are abbreviated as follows: ROUGE-1 (R1), ROUGE-L (RL), METEOR (MET), Precision (P), Recall (R), Normalized Discounted Cumulative Gain (NDCG), Hit Ratio (HR), Accuracy (ACC), Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Perplexity (PPL), and additional metrics for opinion distribution analysis such as Representativeness, Steerability, Consistency, and Wasserstein Distance (Rep/St/Con/Was). "N/A" in the Output column indicates that the target is non-textual, while in the Evaluation Metrics column, it signifies the absence of established metrics for that dataset in personalization tasks, although the dataset may have potential for personalized applications.

	Data	Size (users)	Output	Eval. Metrics	User Text	Numer. Attrs.	Cat. Attrs.	Timestamps	Filtered	
SHORT-TEXT GENERATION	News Headline	13K/2K/2K	9.7	R1/RL	✓			✓	✓	Salemi et al. (2023)
	Scholarly Title	10K/3K/3K	9.2	R1/RL	✓			✓	✓	Salemi et al. (2023)
	Email Title	5K/1K/1K	7.3	R1/RL	✓			✓	✓	Salemi et al. (2023)
	Tweet Paraphrase	10K/2K/1K	16.9	R1/RL	✓			✓	✓	Salemi et al. (2023)
LONG-TEXT GENERATION	Email Generation	3K/1K/1K	92 ± 60	R1/RL, MET.	✓			✓	✓	Kumar et al. (2024)
	Abstract Generation	23K/5K/5K	160 ± 70	R1/RL, MET.	✓	✓	✓	✓	✓	Kumar et al. (2024)
	Review Generation	16K/2K/2K	296 ± 229	R1/RL, MET.	✓	✓	✓	✓	✓	Kumar et al. (2024)
	Topic Writing	16K/2K/2K	262 ± 241	R1/RL, MET.	✓	✓	✓	✓	✓	Kumar et al. (2024)
REC.	MovieLens-1M	6K (3.7K items)	N/A	P/R/NDCG		✓	✓	✓	✓	Harper & Konstan (2015)
	Recipe	2.5K (4.1K items)	N/A	P/R/NDCG		✓	✓	✓	✓	Majumder et al. (2019)
	Amazon	233M (15.2M items)	N/A	P/R/NDCG/HR	✓	✓	✓	✓	✓	Ni et al. (2019)
	Microsoft News	1M (161K items)	N/A	P/R/NDCG/HR			✓	✓		Wu et al. (2020)
	BookCrossing	279K (271K items)	N/A	P/R/NDCG/HR		✓	✓		✓	Ziegler et al. (2005)
	TripAdvisor	10K (6K items)	N/A	P/R/NDCG/HR	✓	✓	✓	✓	✓	Li et al. (2020)
	Yelp	27K (20K items)	N/A	P/R/NDCG/HR	✓	✓	✓	✓	✓	Yelp. (2014)
CLASSIF.	MovieLens-1M	6K (3.7K items)	N/A	P/R/NDCG			✓	✓	✓	Salemi et al. (2023)
	Movie Tag.	4K/0.7K/0.9K	N/A	ACC/F1			✓	✓	✓	Salemi et al. (2023)
	Product Rat.	20K/2.5K/2.5K	N/A	MAE/RMSE			✓	✓	✓	Salemi et al. (2023)
DIALOGUE	ConvAI2	1.2K	13.7	PPL/F1/HR	✓					Dinan et al. (2020)
	Empathetic Conv.	1.9K (79 personas)	N/A	N/A	✓	✓	✓		✓	Omitaomu et al. (2022)
	PRISM	8K	N/A	N/A	✓	✓	✓	✓	✓	Kirk et al. (2024)
QUEST. ANS.	OpinionQA	1.5K	N/A	Rep/St/Con/Was						Santurkar et al. (2023)
	GlobalOpinionQA	2.6K	N/A	Similarity			✓			Durmus et al. (2023)

7 Taxonomy of Datasets for Personalized LLMs

We also categorize the datasets based on whether they are used for direct evaluation of personalized text generation through reward signals that reflect user preferences or for indirect evaluation in a downstream

application. In the former case, as most existing datasets for personalized text generation rely on ground-truth text for evaluation, our discussion primarily focuses on these scenarios. In the latter case, the goal is to demonstrate that incorporating the generated text improves the performance of a downstream task, such as a recommendation model, classifier, or other functional systems.

- **Personalized Datasets *with* Ground-Truth Text (Sec. 7.1)**: Datasets containing actual ground-truth text written by users are relatively rare but essential, as they allow for direct evaluation of personalized text generation methods, rather than relying on performance in downstream tasks. Categories of datasets useful for direct evaluation of personalized text generation include **short-text generation** and **long-text generation** (Figure 5).
- **Personalized Datasets *without* Ground-Truth Text (Sec. 7.2)**: Datasets suited for indirect evaluation via downstream applications are far more common, as they do not require ground-truth text authored by users. These datasets are typically employed to evaluate personalized LLM techniques through tasks such as **recommendation**, **classification**, **dialogue**, and **question answering**.

Table 5 provides a comprehensive summary of various personalization tasks along with their key attributes. For each benchmark dataset, we indicate whether the data has been filtered to include only users with substantial prior activity (e.g., users who have reviewed at least 100 products, sent at least 100 emails, or rated a minimum of k movies), which is relevant for addressing the cold-start problem. We also summarize whether the dataset contains user-written text, numerical attributes (such as ratings), and other categorical attributes, such as the genre of a movie watched. Additionally, we note if the dataset includes text descriptions (e.g., a movie's description) that are not written by the user but may still contribute to personalized LLM techniques, even though they are not user-specific.

7.1 Personalized Datasets *with* Ground-Truth Text

In terms of personalization datasets that actually contain the text written by users which can then be used for direct evaluation of the generated text fall under two main categories in our taxonomy shown in Table 5, notably, **short-text generation** and **long-text generation**. For instance, the review generation benchmark under long-text consists of all the reviews written by a user, the review title for each, and the ratings for every review as well, whereas most of the short-text generation datasets in Table 5 mostly consist of only the title of a news article or the title of an email. Notably, the output length column highlights the difference between **short-text generation** and **long-text generation** tasks. In particular, **personalized short-text generation** data seeks to generate very short text with a few words (*e.g.*, 9-10 words), and is somewhat similar to paraphrasing and summarization, as most of these datasets seek to generate a title of a paper, news article, email, among others. In contrast, data for benchmarking **personalized long-text generation** techniques is significantly much longer and thus more challenging as the goal is to generate a longer piece of text that is often 100s or 1000s of words. Nevertheless, all datasets under either of the proposed categories in our taxonomy, namely, **short-text generation** and **long-text generation**, can be used for directly evaluating the generated text as well as using the text provided for developing personalized LLM techniques via training or fine-tuning LLMs using the user-specific text provided for training.

7.2 Personalized Datasets *without* Ground-Truth Text

Personalized datasets that are useful for indirect evaluation of the generated text via a downstream application are by far the most common, as they do not require an actual set of ground-truth text written by individual users, and instead can leverage user attributes and other interaction data, to generate personalized text, and then this text is used to enhance another arbitrary model such as a recommendation approach. Such datasets that can be used in this fashion are shown in Table 5 in the following categories, including, **recommendation**, **classification**, **dialogue**, and **question answering**. Leveraging such category of evaluation strategy enables one to leverage any commonly used dataset that has been used for a variety of different personalization tasks such as recommendation, classification, and so on. However, a criticism of this approach is that we can only demonstrate that using the text is useful for the downstream application, but not that the text was actually meaningful or relevant to the user. On the other hand, for instance, generated intermediate tokens or embeddings can be used to augment user embeddings. The augmented embeddings are then shown

to improve the performance of downstream tasks, highlighting the potential value of this approach while acknowledging its inability to validate user-specific relevance directly.

Additionally, we highlight datasets that, while not originally designed for personalization tasks, contain rich user-specific information and user-generated text, making them suitable for downstream personalization tasks. Notable examples include the PRISM Alignment Dataset (Kirk et al., 2024), which links survey responses and demographic information from diverse participants to their interactions with various LLMs, and the Empathic Conversations dataset (Omitaomu et al., 2022), which features multi-turn conversations enriched with detailed demographic and personality profiles, pre- and post-conversation data, and turn-level annotations. These datasets are valuable for tasks such as personalized QA, dialogue generation, empathy-aware responses, and user-specific emotion modeling in language models. We believe that although these datasets do not primarily use user-generated text as ground truth for evaluation, their rich user-specific information has the potential to capture diverse preferences. Developing effective reward models based on this information could be valuable for evaluating personalized text generation in cases where ground-truth user-generated text is unavailable, an area that remains largely unexplored.

8 Applications of Personalized LLMs

In this section, we explore various use cases where personalized LLMs have shown significant potential for enhancing user experiences and improving outcomes.

8.1 Personalized AI Assistant

8.1.1 Education

Personalized LLMs show promising potential to facilitate personalized education experiences for both students and teachers (Gonzalez et al., 2023; Wang et al., 2024b; Jeon & Lee, 2023; Huber et al., 2024; Wang & Demszky, 2023; Joshi et al., 2023; Yan et al., 2024b; Leong et al., 2024) and there have been increasing number of works (Sharma et al., 2023; Elkins et al., 2023; Ochieng, 2023; Olga et al., 2023; Phung et al., 2023) with such ideas. For example, they can analyze students' writing and responses, providing tailored feedback and suggesting materials that align with the students' specific learning needs (Kasneci et al., 2023). EduChat (Dan et al., 2023) tailors LLMs for educational applications by pre-training on educational corpora and stimulating various skills with tool use and retrieval modules through instruction tuning. It offers customized support for tasks such as Socratic teaching, emotional counseling, and essay assessment. Tu et al. (2023) use ChatGPT to create educational chatbots for teaching social media literacy and investigate ChatGPT's ability to pursue multiple interconnected learning objectives, adapt educational activities to users' characteristics (such as culture, age, and education level), and employ diverse educational strategies and conversational styles. While ChatGPT shows certain ability to adapt educational activities based on user characteristics with diverse educational strategies, the study identifies challenges such as the limited conversation history, highly structured responses, and variability in ChatGPT's output, which can sometimes lead to unexpected shifts in the chatbot's role from teacher to therapist. Park et al. (2024b) present a personalized tutoring system that leverages LLMs and cognitive diagnostic modeling to provide tailored instruction in English writing concepts. The system incorporates student assessment across cognitive state, affective state, and learning style to inform adaptive exercise selection and personalized tutoring strategies implemented through prompt engineering. While their proposed system shows promise in adapting to individual students, the authors identified challenges in translating assessments into effective strategies and maintaining engagement, pointing to areas for further research in LLM-based personalized education. Overall, the challenges of personalized LLMs in education include copyright and plagiarism issues, biases in model outputs, over-reliance by students and teachers, data privacy and security concerns, developing appropriate user interfaces, and ensuring fair access across languages and socioeconomic backgrounds (Kasneci et al., 2023).

8.1.2 Healthcare

LLMs have demonstrated remarkable proficiency in various healthcare-related tasks (Liu et al., 2023e; Wang et al., 2023c; Liu et al., 2023h; Yang et al., 2024b), paving the way for their potential integration into

personalized health assistance. Belyaeva et al. (2023) introduce HeLM, a framework that enables LLMs to leverage individual-specific multimodal health data for personalized disease risk prediction. HeLM employs separate encoders to map non-text data modalities (such as tabular clinical features and high-dimensional lung function measures) into the LLM's token embedding space, allowing the model to process multimodal inputs together. Abbasian et al. (2023) present openCHA, an open-source LLM-powered framework for conversational health agents that enables personalized healthcare responses by integrating external data sources, knowledge bases, and analytical tools. The framework features an orchestrator for planning and executing information-gathering actions and incorporates multimodal and multilingual capabilities. Building on openCHA, Abbasian et al. (2024) integrate specific diabetes-related knowledge to enhance performance in downstream tasks within the domain. Zhang et al. (2024a) introduce MaLP, a novel framework for personalizing LLMs as medical assistants. The approach combines a dual-process enhanced memory (DPeM) mechanism, inspired by human memory processes, with PEFT to improve LLMs' ability to provide personalized responses while maintaining low resource consumption. Jin et al. (2024b) propose a Health-LLM-based pipeline with RAG to provide personalized disease prediction and health recommendations. The system extracts features from patient health reports using in-context learning, assigns scores to these features using medical knowledge, and then employs XGBoost for final disease prediction.

8.1.3 Other Domains

Beyond the two domains where personalized LLMs have been widely employed, this section explores areas with less focus but significant potential for applying personalized LLMs. In these domains, specialized LLMs or language agent frameworks are emerging, but they often lack emphasis on personalization—an aspect that could greatly enhance user experiences.

Finance Beyond the general advances of LLMs in finance (Araci, 2019; Wu et al., 2023b), personalized LLMs have shown significant potential in providing tailored financial advice, extending beyond general investment recommendations. For instance, Liu et al. (2023d) introduce FinGPT, a model that offers personalized financial advice by considering individual user preferences such as risk tolerance and financial goals. Additionally, Lakkaraju et al. (2023) evaluate the performance of LLMs as financial advisors by posing 13 questions related to personal finance decisions. The study highlights that while these LLM-based chatbots generate fluent and plausible responses, they still face critical challenges, including difficulties in performing numeric reasoning, a lack of visual aids, limited support for diverse languages, and the need for evaluation across a broader range of user backgrounds. Future applications of personalized LLMs in the financial domain could encompass a variety of specialized services. These may include personalized wealth management strategies, where LLMs offer dynamic advice on asset allocation and retirement planning, tailored risk assessment tools that provide custom risk profiles and real-time monitoring, and tax optimization strategies that help individuals and businesses minimize tax liabilities. Furthermore, LLMs could be deployed in personalized insurance solutions, credit management (including tailored loan recommendations and credit score optimization), and spending and budgeting tools that adapt to an individual's financial habits and goals. These applications could significantly enhance the relevance and utility of personalized LLMs in the financial sector.

Legal A growing number of LLMs (Nguyen, 2023; Huang et al., 2023c; Cui et al., 2023) has been developed specifically for legal applications, where these models have proven useful in assisting judges with decision-making, simplifying judicial procedures, and improving overall judicial efficiency (Lai et al., 2023; Trautmann et al., 2022; Blair-Stanek et al., 2023; Yu et al., 2022; Nay, 2023; Fei et al., 2024). DISC-LawLLM (Yue et al., 2023) fine-tunes LLMs using legal syllogism prompting strategies and enhances them with a retrieval module to offer a broad range of legal services, potentially leveraging personal historical data. He et al. (2024b) introduce SimuCourt, a judicial benchmark comprising 420 real-world Chinese court cases, to evaluate AI agents' judicial analysis and decision-making capabilities. SimuCourt integrates AgentsCourt, a novel multi-agent framework that simulates court debates, retrieves legal information, and refines judgments using LLMs. This framework allows for the integration of different personas into various agents, enabling personalized interactions throughout the legal process. Looking ahead, we anticipate that personalized LLMs will significantly assist legal professionals by catering to their specific needs: For **lawyers**, personalized LLMs

can be used for personalized case analysis, where they analyze legal cases in light of a lawyer's past cases, preferences, and typical strategies. This can lead to more effective argumentation and strategy formulation tailored to the lawyer's style. Moreover, personalized LLMs can enhance client interactions by adapting communication styles, content, and language to meet the unique needs of each client. This not only improves client satisfaction but also helps in maintaining long-term client relationships. Additionally, LLMs can assist in drafting legal documents, such as contracts and agreements, by incorporating specific clauses, language, and legal standards preferred by the lawyer or their firm. For **judges**, personalized LLMs can provide support in managing their caseload and ensuring consistent and fair rulings. Specifically, they can generate personalized case summaries that highlight the most relevant details based on a judge's past rulings and areas of focus, such as statutory interpretation or case law precedence. Furthermore, personalized LLMs can offer custom verdict recommendations that align with a judge's legal principles and prior decisions, promoting consistency in judicial outcomes. For **clients**, personalized LLMs can make legal services more accessible and tailored to individual needs. These models can offer personalized legal consultation by analyzing the client's specific circumstances, legal history, and goals, providing advice that is both relevant and easy to understand. Personalized LLMs can also provide clients with regular updates on their cases, offering clear explanations and progress reports that keep them informed and involved in the legal process. In summary, personalized LLMs have the potential to transform the legal domain by providing tailored support to different roles, enhancing efficiency, accuracy, and client satisfaction across the board.

Coding As LLMs continue to advance, especially those fine-tuned on code-specific datasets (Roziere et al., 2023; Chen et al., 2021), their ability to generate high-quality code has seen significant improvement. This has spurred the development of an increasing number of AI-powered assistants aimed at enhancing the coding experience for programmers (Zhang et al., 2024b; Wang et al., 2024e;d; Xia et al., 2024). However, these applications often overlook the crucial aspect of personalization. For personalized code generation, Dai et al. (2024) propose MPCODER a novel approach for generating personalized code for multiple users that aligns with their individual coding styles. This method uses explicit coding style residual learning to capture syntax standards and implicit style learning to capture semantic conventions, along with a multi-user style adapter to differentiate between users through contrastive learning. The authors introduce a new evaluation metric called Coding Style Score to quantitatively assess coding style similarities. Personalization in coding assistance can be realized in several ways. Firstly, programmers and teams often have unique coding styles. For example, a middle-school student might prefer code that is easy to understand and well-commented, while a software engineer in a tech company may prioritize performance, scalability, and strict adherence to industry standards. A personalized LLM can learn from a user's behavior over time and adapt their suggestions to match the user's skill level, preferred frameworks, and commonly used coding patterns would significantly enhance their utility. Secondly, context-aware debugging is another area where personalization can make a substantial impact. Personalized LLMs could offer tailored debugging assistance based on a programmer's typical errors and preferred debugging strategies. Thirdly, enforcing code guidelines that align with a team's standards such as naming conventions, architectural patterns, and tool integrations is essential in collaborative environments. This ensures consistency and maintainability across the codebase, which is particularly critical in professional settings. Finally, personalized LLMs could greatly improve collaboration and code review processes by offering suggestions that consider both individual and team preferences. Achieving such levels of personalization requires advanced techniques like RAG or fine-tuning on user-specific data, enabling the model to adapt to the distinct needs and preferences of different programmers. This represents a promising direction for future research and development in AI-powered coding assistants.

8.2 Recommendation

Personalized LLMs have been extensively applied across various recommendation tasks, including direct recommendations, sequential recommendations, conversational recommendations, explainable recommendations, rating predictions, and ranking predictions (Dai et al., 2023; Du et al., 2024; Hou et al., 2024; Liu et al., 2023a; 2024d; Ji et al., 2024b). These applications aim to enhance user experiences in recommendation domains like e-commerce (Tan & Jiang, 2023; Chen et al., 2024c). According to Wu et al. (2023a), the integration of LLMs into recommendation systems can be categorized into three main approaches: (1) augmenting traditional recommendation systems, such as collaborative filtering (Schafer et al., 2007; Resnick et al.,

1994), with LLM embeddings; (2) enriching traditional recommendation systems by using LLM-generated outputs as features; and (3) employing LLMs directly as recommenders in downstream recommendation tasks. These three categories of approaches can all benefit from the personalization techniques discussed in Sec. 5. To enhance traditional recommendation systems with personalized LLMs, PALR (Yang et al., 2023) leverages LLMs to generate natural language user profiles, retrieve relevant item candidates, and rank and recommend items using a fine-tuned LLM. Chat-Rec improves traditional recommender systems by using LLMs to increase interactivity and explainability. It converts user profiles and historical interactions into prompts, allowing LLMs to learn user preferences through in-context learning and generate more personalized recommendations. For direct use of personalized LLMs as recommenders, InstructRec (Zhang et al., 2023a) frames recommendation tasks as instruction-following tasks for LLMs. It designs a flexible instruction format that incorporates user preferences, intentions, and task forms, and generates a large dataset of personalized instructions to fine-tune an LLM for recommendation tasks. Similarly, GeneRec (Wang et al., 2023b) employs LLMs for personalized content creation based on user instructions and feedback, aiming to complement traditional retrieval-based recommendation systems. While personalized LLMs have seen widespread application in recommendation systems, demonstrating superior performance in few-shot and zero-shot settings with enhanced explainability that addresses the cold-start problem (Liu et al., 2023a; Dai et al., 2023; Hou et al., 2024), significant challenges remain, including concerns over privacy, cost, and latency in large-scale deployments, highlighting the need for continued research and innovation

8.3 Search

Recently, with their growing capabilities in summarization and instruction following, LLMs have been incorporated into search engines (applications such as the new Bing (Microsoft, 2023) and SearchGPT (OpenAI, 2024)), which can provide an engaging conversational process that can help users find information more effectively (Spatharioti et al., 2023; Joko et al., 2024). Incorporating personalization can further tailor results to individual users' search histories, interests, and contexts, which can lead to more relevant and efficient search experiences (Bennett et al., 2012; Harvey et al., 2013; Cai et al., 2014; Song et al., 2014; Vu et al., 2014; 2015; Zhou et al., 2021). A large number of works (Dou et al., 2007; Sieg et al., 2007; Carman et al., 2010; Teevan et al., 2011; White et al., 2013) focused on how to better personalize search engines before the emergence of LLMs. In the era of LLMs, Zhou et al. (2024b) propose Cognitive Personalized Search (CoPS), a personalized search model that combines LLMs with a cognitive memory mechanism inspired by human cognition. CoPS utilizes sensory, working, and long-term memory components to efficiently process user interactions and improve search personalization without requiring training data. Besides, Jiang et al. (2024b) introduce Collaborative STORM, a system that personalizes search experiences by engaging users in multi-turn search sessions and incorporating their interaction history. However, despite the advancements in LLM-augmented search engines, there are still challenges to be addressed. Notably, Sharma et al. (2024) find that users of LLM-powered search exhibit more biased information querying and opinion polarization compared to conventional web search, especially when the LLM reinforces existing views. This phenomenon, known as the "echo chamber" effect (Pariser, 2011; Sharma et al., 2024; Lazovich, 2023; Garimella et al., 2018), emphasizes the challenges in balancing personalization with the need for diverse and objective information retrieval.

9 Open Problems & Challenges

Despite the significant progress made in the applications of personalized LLMs, there remain numerous unresolved challenges and open research questions. In this section, we explore key issues that require further investigation and innovation to advance the field. These challenges span various aspects of personalization, including the development of reliable benchmarks and evaluation metrics, tackling the persistent cold-start problem, addressing concerns about stereotypes and bias in personalized models, ensuring privacy in user-specific data handling, and expanding personalization to multi-modal systems. Each of these areas presents unique challenges that must be overcome to achieve more robust, fair, and effective personalized LLMs.

9.1 Comprehensive Evaluation: Benchmarks, Automatic Metrics, LLM-as-a-Judge, and Beyond

Effective benchmarks, combined with comprehensive metrics, are crucial for evaluating various aspects of LLMs, including their ability to personalize outputs. However, existing benchmarks for personalization are largely derived from recommendation systems, where the focus is predominantly on final predictions such as ratings, recommended items, or rankings. These benchmarks often overlook the intermediate processes in LLMs' output generation, which are critical for assessing whether the output is genuinely personalized. LaMP (Salemi et al., 2023) is one of the few benchmarks that specifically targets the evaluation of LLMs in generating personalized outputs. However, LaMP's scope is limited to text classification and short, single-turn text generation tasks. It lacks the complexity of real-world interactions, which are essential for applications like personalized AI assistants. This gap highlights the need for new benchmarks that can evaluate LLMs' personalized output generation in more realistic scenarios. Such benchmarks should also integrate personalization perspectives into other key LLM capabilities, including reasoning, planning, instruction following, and long-context understanding, thereby providing a more holistic evaluation. Overall, we envision comprehensive personalization benchmarks that effectively capture multi-turn interactions, evolving user preferences, and diverse contextual scenarios. For future directions in designing personalization benchmarks, one promising area involves language agents, which are increasingly applied to tasks such as web navigation (Deng et al., 2024), scientific discovery (Chen et al., 2024e; Si et al., 2024), research support (Asai et al., 2024), and coding assistance (Wang et al., 2024e). While their evaluation typically focuses on task success rates, the role of personalization in these contexts remains underexplored and calls for the development of dedicated evaluation criteria. Additionally, while current benchmarks predominantly emphasize English, there is an urgent need to expand their scope to include multilingual settings to ensure broader applicability and inclusivity. Given that LLMs are known to be sensitive to prompt variations (Zhuo et al., 2024), addressing distribution shifts and robustness in prompt design, as well as out-of-distribution scenarios (Wang et al., 2023a), becomes essential. Controlled experiments that systematically isolate and analyze dimensions such as style, topic, and user preferences can provide valuable insights. Furthermore, incorporating cultural and values-based adaptations (Shi et al., 2024b), alongside dialectal variations (Ziems et al., 2023), will enable these benchmarks to better reflect the real-world complexity of personalization and its diverse user base. Additionally, evaluating on static benchmarks often entails limitations, such as susceptibility to data contamination (Sainz et al., 2023; Balloccu et al., 2024) and a lack of adaptability to evolving LLM capabilities. To address these issues, designing dynamic evaluation frameworks for personalization tasks offers a promising alternative, enabling assessments with controllable complexity (Zhu et al., 2023b; Zhang et al., 2024c).

In addition, there is currently no comprehensive quantitative metric to assess the degree of personalization in LLM-generated outputs. Most existing metrics are task-specific and heavily dependent on the downstream task formulations and the quality of gold labels. As a result, they often fail to capture the diverse dimensions of personalization, such as those illustrated in Figure 4. The recent trend of using LLMs as judges in evaluating various aspects of LLM-generated content, due to their versatile nature, presents a promising approach for personalization assessment. Designing an LLM-as-a-Judge framework with personalized criteria rubrics could offer a more nuanced evaluation of the degree of personalization in LLM outputs. However, this approach remains underexplored, and challenges such as instability and potential biases need to be addressed to make it a reliable evaluation method. Specifically, challenges arise when LLM-based evaluators unintentionally inject their own internal persona or biases (Jiang et al., 2024a), and they can sometimes exhibit overconfidence by offering overly favorable assessments of their own outputs (Tian et al., 2023). Consistency is another issue (Li et al., 2024c), as LLM-as-a-judge evaluations often fluctuate depending on the context (Gu et al., 2024) and exhibit phenomena such as positional biases (Li et al., 2024i), which poses significant problems in personalization tasks involving nuanced user attributes. There are also concerns about robustness, given that adversarial inputs can undermine the integrity of LLM-based evaluations (Doddapaneni et al., 2024; Raina et al., 2024; Shi et al., 2024a; Zheng et al., 2024). Furthermore, efficiency and flexibility can be limited when LLM-based evaluators must be manually prompted for each new personalization scenario. Despite these challenges, there are significant opportunities for improvement. Developing more refined prompting formats and strategies, such as scalar scoring, pairwise comparisons, and multiple-choice selections, can provide more nuanced and reliable evaluations. Additionally, addressing open problems like in-context exemplar selection could further enhance the evaluation process. Given the computational overhead of large-scale LLM-as-a-judge frameworks, which often rely on extensive API calls for model inference (Gu et al., 2024),

training specialized smaller models tailored for personalization evaluation presents a promising alternative for reducing costs and improving efficiency (Kim et al., 2024; Huang et al., 2024a). Furthermore, exploring agent-as-a-judge paradigms (Zhuge et al., 2024) that incorporate multi-agent collaboration, tool integration, or human-in-the-loop approaches offers a path toward greater transparency, fairness, and robustness in assessing personalized outputs.

In summary, evaluating robust personalization in LLMs presents unique challenges compared to the evaluation of other capabilities. Personalization objectives in real-life applications are inherently diverse, leading to pluralistic alignment requirements (Sorensen et al., 2024). The varying levels of personalization defy reliance on any single evaluation criterion. Furthermore, the inherent ambiguity and subjectivity of user preferences add complexity, underscoring the urgent need for a finer-grained taxonomy to capture the full spectrum of personalization phenomena. Current datasets often suffer from imbalanced data, which can lead to incomplete or biased assessments of personalization performance. For instance, datasets may overrepresent users with frequent online activity, potentially skewing models to prioritize the preferences and behaviors of these users while underrepresenting less active user groups. Additionally, user preferences are dynamic and can shift over time, necessitating the development of multi-turn benchmarks that reflect evolving personalization goals and longitudinal variations. Moreover, due to inconsistencies in dataset structures, task formulations, and the diverse scenarios in real-life applications, direct performance comparisons can be challenging, making it difficult to achieve a unified and fair evaluation of different techniques. It is also critical to address privacy, bias, and fairness—factors that may conflict with certain personalization objectives—to ensure personalized systems remain safe, ethical, and equitable. Given these challenges, we advocate for collaborative efforts to create open-source platforms and shared tasks that allow personalized LLMs to be tested against standardized metrics aligned with core LLM capabilities. Finally, we emphasize the importance of cross-domain generalization tests beyond recommendation systems and text generation to evaluate how well personalization systems adapt across diverse domains and contexts.

9.2 Cold-start Problem

The cold-start issue is a prevalent and challenging problem in recommendation systems, where the system must generate recommendations for items that have not yet been rated by any users in the dataset, or when there is minimal information available about user preferences (Schein et al., 2002; Guo, 1997). Previously, a large number of methods (Lam et al., 2008; Li et al., 2019; Park & Chu, 2009; Lee et al., 2019; Wei et al., 2017) have been proposed to address such issues in traditional recommendation systems. Although LLMs demonstrate strong few-shot capabilities through in-context learning and role-playing via instructional prompting, significant challenges remain in effectively adapting personalized LLMs to sparse user-specific data via fine-tuning. This issue is further compounded by the fact that many downstream datasets are preprocessed to exclude instances with limited user interaction history—often filtering out data points where fewer than five interactions are recorded. As a result, the potential of personalized LLMs to handle low-resource scenarios remains relatively underexplored, and more advanced techniques are required to improve their adaptation to sparse data settings. Persona-DB (Sun et al., 2024) addresses cold-start problems more effectively through a hierarchical construction process that distills abstract, generalizable personas from limited user data. This is followed by a collaborative refinement stage that leverages similarities between users to fill knowledge gaps, allowing the system to draw relevant insights from users with richer interaction histories when personalizing for new or infrequent users. Given the limited number of works on personalized LLMs, we propose two potential research directions: (1) Building on the Persona-DB paradigm, through stages of prompting and abstraction that progressively refine and generalize user personas, potentially improving personalization across diverse applications. (2) Leveraging synthetic data generation, which has shown promise in enhancing various LLM capabilities (Chan et al., 2024; Zhang et al., 2024c; Tong et al., 2024). LLMs could be employed to generate large-scale user-specific data from sparse seed data. However, challenges such as ensuring diversity and maintaining high-quality data at scale remain key obstacles to this approach.

9.3 Stereotype & Bias Issues of Personalization

The personalization of LLMs introduces significant concerns regarding the amplification and perpetuation of stereotypes and biases (Zhang et al., 2023d; Ziems et al., 2024; Gallegos et al., 2024; Li et al., 2023f).

When LLMs generate personalized outputs, they rely on data that may inherently contain societal biases related to gender, race, ethnicity, culture, and other sensitive attributes. Personalization can unintentionally reinforce these biases by tailoring content that aligns with the biased data the models are trained on or the ones provided in the prompt, thus exacerbating the problem. For example, recent research (Gupta et al., 2023; Deshpande et al., 2023; Cheng et al., 2023a; Wan et al., 2023) indicates that assigning a specific persona to LLMs, like that of a "disabled person," can unexpectedly alter their performance across diverse tasks, including those seemingly unrelated to the assigned identity such as general knowledge or mathematical reasoning. Moreover, the feedback loop created by personalized systems can further entrench biases. Besides, Weissburg et al. (2024) examines bias in LLMs as personalized educators, showing disparities in content selection and difficulty based on demographics like race, gender, income, and disability. Using 17,000+ explanations and two bias metrics, it finds all tested LLMs exhibit bias, with income and disability status most affected. As LLMs continue to adapt based on user interactions, they might cater to pre-existing preferences and viewpoints, reducing the opportunity for exposure to diverse or corrective perspectives. This can lead to the deepening of echo chambers, where users are repeatedly exposed to biased or stereotypical information without opportunities for counterbalance. For example, Kantharuban et al. (2024) investigates how large language models generate recommendations that reflect both explicit and implicit cues of a user's identity, leading to biased outputs that align with racial stereotypes. The authors show that although personalization can tailor recommendations to the user's background, it often restricts the range of options for minority users and obscures the role of identity in the recommendation process. Despite growing efforts to mitigate biases in LLMs (Lu et al., 2020; Han et al., 2021; Ghanbarzadeh et al., 2023; Zhang et al., 2023d), there is a limited number of works on how personalization intersects with these biases. He et al. (2024a) introduce Context Steering (CoS), a training-free method for enhancing personalization and mitigating bias in LLMs at inference time by quantifying and modulating the influence of contextual information on model outputs. CoS works by computing the difference in token prediction likelihoods between models with and without context, then using this difference to adjust the influence of context during generation. However, CoS does not specifically aim to mitigate biases that arise within the personalization process. Vijjini et al. (2024) introduce the concept of personalization bias in LLMs, where LLM performance varies based on the user's demographic identity provided during the interaction. It proposes a framework to evaluate and quantify this bias by measuring safety-utility trade-offs across different user identities, using widely-used datasets. The study demonstrates the prevalence of personalization bias in both open-source and closed-source LLMs, and while it explores mitigation strategies like preference tuning and prompt-based defenses, it concludes that completely eliminating this bias remains an open challenge, highlighting the need for further research in this area. Overall, it is critical to design personalization systems that actively account for fairness and inclusivity, ensuring that personalized outputs do not reinforce harmful stereotypes or perpetuate biased perspectives. Future work should explore techniques such as bias detection during the personalization process, incorporating fairness constraints into the personalization pipeline, and ensuring that diverse perspectives are represented in user-specific outputs. For example, general LLM debiasing methods can serve as inspiration, including data-centric approaches such as data augmentation, filtering, reweighting training data, or modifying loss functions during fine-tuning to directly mitigate biases. Additionally, inference-time techniques—such as adjusting decoding strategies or rewriting model outputs—can help reduce bias in personalized responses while maintaining the model's tailored behavior (Gallegos et al., 2024). Addressing these challenges will require careful consideration of the trade-offs between personalization and fairness, as well as the development of robust evaluation frameworks that measure not only the degree of personalization but also the impact on bias and stereotype propagation. Ultimately, ensuring that personalized LLMs promote inclusivity and fairness is essential to building systems that are socially responsible and beneficial to all users.

9.4 Privacy Issues

Privacy, particularly concerning Personally Identifiable Information (PII), is a critical concern in LLM personalization applications, where the objectives of personalization and privacy often conflict. Current LLMs are vulnerable to privacy breaches, as they can accurately infer personal attributes (e.g., location, income, gender) from unstructured text, even when common mitigations such as text anonymization and model alignment are employed (Staab et al., 2023). Additionally, adversarial attacks, such as prompt injections (Liu et al., 2023f; Zhan et al., 2024; Ning et al., 2024a) and jailbreaking (Xu et al., 2024b; Wei et al., 2024a),

can cause LLMs to generate inappropriate content or reveal sensitive information from their training data. Although a growing body of research focuses on addressing privacy leakage in LLMs (Behnia et al., 2022; Lukas et al., 2023; Chen et al., 2023b; Yao et al., 2024; Yan et al., 2024a; Kuang et al., 2024; Feng et al., 2024), there is limited work specifically targeting the intersection of personalization and privacy (Zhao et al., 2024a). To address this gap, it is crucial to formally define the boundary between personalization and privacy, which may vary across tasks and can be subjective for different users. Furthermore, it is essential to design specialized modules that prevent both explicit and implicit privacy leaks throughout various stages of LLM personalization, such as data processing, model training, and retrieval processes. An ideal solution would allow for flexible adjustment, enabling a balanced trade-off between the degree of personalization and privacy protection, tailored to individual user preferences and specific application contexts.

9.5 Multimodality

Personalizing large multimodal models (LMMs) is particularly complex due to the diverse nature of the data they process, such as text, images, audio, and video. These models are designed to handle multiple input types and fuse them in meaningful ways to improve task performance across various domains. However, personalization in this context introduces unique challenges as it must account for individual user preferences or characteristics across multiple modalities simultaneously. In personalized multimodal models, the key challenge lies in effectively integrating user-specific data, such as preferences or interaction history, to modulate the model's responses in a contextually appropriate manner. These user-specific data points may come from different modalities as well. For example, in personalized image generation tasks, the model must be able to generate images that align with user-specific visual preferences while also understanding the associated textual or auditory cues. Recent work, such as Unified Multimodal Personalization (Wei et al., 2024b), demonstrates how LMMs can be adapted for personalization across multiple tasks and modalities. This framework unifies the personalization of tasks like preference prediction, explanation generation, and image generation under a common structure, leveraging the strengths of multimodal learning to predict user-specific outcomes based on a combination of text, images, and potentially other input types. Another notable example is the work on multi-subject personalization for text-to-image models (Jang et al., 2024), which focuses on personalizing generated images to represent distinct user preferences for multiple subjects within a single image. Similarly, techniques like personalized multimodal generative models (Shen et al., 2024) have been proposed to handle multiple modalities by transforming user behavior data into natural language for further personalization, extending the utility of LLMs in multimodal scenarios. Personalized multimodal models also pose significant computational challenges. The fusion of modalities requires more sophisticated architectures capable of jointly learning from multiple data streams without sacrificing personalization fidelity. For instance, plugging visual embeddings into traditional recommendation systems and personalized image generation models leverage LMMs to enhance the personalization process by embedding user preferences from both textual and visual input. To further push the boundaries of multimodal personalization, integrating modalities like video or audio into user-centric applications like recommendation systems and content generation presents another layer of complexity. Handling synchronization between modalities, ensuring cohesive user representations, and balancing personalization across dynamic content remain open challenges.

Specifically, according to Wu et al. (2024a), the unique high-level challenges in personalized multimodal LLM systems can be categorized as follows. (1) The integration of heterogeneous data. LMMs must process and align diverse data types such as text, images, audio, and structured user interaction history (Li et al., 2024f). Encoding discrepancies arise as different modalities require distinct encoding techniques, making fusion difficult. Additionally, cross-modal alignment remains an issue since text and images may contain complementary but sometimes conflicting information. Another difficulty is data sparsity, where some users interact predominantly with one modality, leading to imbalanced personalization. Future opportunities include developing unified embedding spaces to bridge the gap between diverse modalities, leveraging self-supervised learning to enhance cross-modal understanding, and designing adaptive models that dynamically adjust weights for each modality based on user interaction patterns. (2) Data noise, redundancy, and quality control. Different modalities often include noisy, redundant, or irrelevant information (Liu et al., 2024e; Lyu & Luo, 2022). For example, images of the same object may vary in quality, while textual descriptions may be verbose or contain unnecessary details. Extracting meaningful insights while filtering out redundant or noisy data is essential for effective personalization. Future work should focus on implementing noise-aware training

strategies to filter out irrelevant data during model training, using multimodal attention mechanisms to prioritize relevant user interactions and suppress redundant information, and employing contrastive learning to improve robustness against low-quality or inconsistent data. (3) Granular understanding of multimodal data. While text-based LLMs excel at linguistic processing, capturing subtle visual or auditory cues remains difficult (Shen et al., 2024). User preferences in areas such as fashion, art, or music often depend on fine-grained multimodal details, such as color, texture, or rhythm. Personalized multimodal models must enhance their ability to extract and relate these details meaningfully across different input types. Future advancements should develop hierarchical representations that preserve both fine-grained and high-level information, improve multimodal contrastive learning to align visual, auditory, and textual representations effectively, and explore fine-tuned retrieval-based techniques to improve nuanced personalization.

Beyond these overarching challenges, specific tasks introduce unique obstacles. In personalized image generation, achieving a balance between fidelity and diversity remains a key challenge. Hybrid diffusion models have shown promise in addressing this issue by refining subject representation and controlling generation constraints (Wang et al., 2024c; Ma et al., 2024; Song et al., 2025). Additionally, tokenization presents another challenge, as indexing multimodal inputs into lookup tables for efficient personalized generation requires specialized designs (Gal et al., 2022). For multi-modal personalized recommendation, multi-modal collaborative filtering introduces additional complexities. Recent approaches have attempted to unify multi-channel information, integrating generative recommendation mechanisms alongside dynamic item modifications (Yu et al., 2024). These methods enable tokenization of both items and user embeddings, facilitating multimodal feature incorporation into the model's latent space. Addressing these challenges will require continued research into improved multimodal alignment, data filtering, and scalable architectures that enhance personalization across diverse applications. Future advancements in integrating multimodal feedback loops and adaptive learning mechanisms will be essential for unlocking the full potential of personalized multimodal LLMs.

10 Conclusion

This survey provides a unified and comprehensive view of the burgeoning field of personalized LLMs. It bridges the gap between the two dominant lines of research — direct personalized text generation and leveraging personalized LLMs for downstream applications — by proposing a novel taxonomy for personalized LLM usage and formalizing their theoretical foundations. An in-depth analysis of the personalization granularity highlights trade-offs among user-level, persona-level, and global preference alignment approaches, laying the groundwork for future hybrid systems that can dynamically adapt to user needs and data availability. Additionally, this survey provides a detailed examination of techniques for personalizing LLMs, shedding light on the strengths and limitations of each approach. We explore the nuances of retrieval-augmented generation, various prompting strategies, representation learning approaches, and the evolving landscape of learning from personalized feedback through RLHF, underscoring the need for more robust and nuanced methods. Finally, our comprehensive survey of evaluation methodologies and datasets highlights the critical need for new benchmarks and metrics specifically designed for assessing personalized LLM outputs. Despite the significant progress made in personalized LLMs, numerous challenges and open problems remain. Key areas for future research include addressing the cold-start problem in low-resource scenarios, mitigating stereotypes and biases in personalized outputs, ensuring user privacy throughout the personalization pipeline, and extending personalization to multi-modal systems. The field of personalized LLMs is rapidly evolving, with the potential to revolutionize human-AI interaction across diverse domains. By understanding the foundations, techniques, and challenges outlined in this survey, researchers and practitioners can contribute to the development of more effective, fair, and socially responsible personalized LLMs that cater to diverse user needs and preferences.

References

Mahyar Abbasian, Iman Azimi, Amir M Rahmani, and Ramesh Jain. Conversational health agents: A personalized llm-powered agent framework. *arXiv preprint arXiv:2310.02374*, 2023.

Mahyar Abbasian, Zhongqi Yang, Elahe Khatibi, Pengfei Zhang, Nitish Nagesh, Iman Azimi, Ramesh Jain, and Amir M Rahmani. Knowledge-infused llm-powered conversational health agent: A case study for diabetes patients. *arXiv preprint arXiv:2402.10153*, 2024.

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

Gati V Aher, Rosa I. Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 337–371. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/aher23a.html.

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms. *arXiv preprint arXiv:2402.14740*, 2024.

Amazon. How to use amazon rufus, 2024. URL https://www.aboutamazon.com/news/retail/how-to-use-amazon-rufus. Accessed: 2024-09-18.

Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, et al. Foundational challenges in assuring alignment and safety of large language models. *arXiv preprint arXiv:2404.09932*, 2024.

Negar Arabzadeh, Xinyi Yan, and Charles L. A. Clarke. Predicting efficiency/effectiveness trade-offs for dense vs. sparse retrieval strategy selection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, pp. 2862–2866, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384469. doi: 10.1145/3459637.3482159. URL https://doi.org/10.1145/3459637.3482159.

Dogu Araci. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*, 2019.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023. doi: 10.1017/pan.2023.2.

Akari Asai, Sewon Min, Zexuan Zhong, and Danqi Chen. Retrieval-based language models and applications. In Yun-Nung (Vivian) Chen, Margot Margot, and Siva Reddy (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 6: Tutorial Abstracts)*, pp. 41–46, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-tutorials.6. URL https://aclanthology.org/2023.acl-tutorials.6/.

Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D'arcy, et al. Openscholar: Synthesizing scientific literature with retrieval-augmented lms. *arXiv preprint arXiv:2411.14199*, 2024.

AutoGPT. AutoGPT, 2024. URL https://github.com/Significant-Gravitas/AutoGPT. Accessed: 2024-09-18.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. In Yvette Graham and Matthew Purver (eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 67–93, St. Julian's, Malta, March 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.eacl-long.5/.

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

Keqin Bao, Jizhi Zhang, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. Tallrec: An effective and efficient tuning framework to align large language model with recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pp. 1007–1014, 2023.

Murray R Barrick and Michael K Mount. The big five personality dimensions and job performance: a meta-analysis. *Personnel psychology*, 44(1):1–26, 1991.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, et al. Llms instead of human judges? a large scale empirical study across 20 nlp evaluation tasks. *arXiv preprint arXiv:2406.18403*, 2024.

Jonathan Baxter and Peter Bartlett. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15:319–350, 2001.

Rouzbeh Behnia, Mohammadreza Reza Ebrahimi, Jason Pacheco, and Balaji Padmanabhan. Ew-tune: A framework for privately fine-tuning large language models with differential privacy. In *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 560–566. IEEE, 2022.

Richard Bellman. *Dynamic Programming.* Princeton University Press, Princeton, NJ, 1957.

Anastasiya Belyaeva, Justin Cosentino, Farhad Hormozdiari, Krish Eswaran, Shravya Shetty, Greg Corrado, Andrew Carroll, Cory Y McLean, and Nicholas A Furlotte. Multimodal llms for health grounded in individual-specific data. In *Workshop on Machine Learning for Multimodal Healthcare Data*, pp. 86–102. Springer, 2023.

Paul N Bennett, Ryen W White, Wei Chu, Susan T Dumais, Peter Bailey, Fedor Borisyuk, and Xiaoyuan Cui. Modeling the impact of short-and long-term behavior on search personalization. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pp. 185–194, 2012.

Andrew Blair-Stanek, Nils Holzenberger, and Benjamin Van Durme. Can gpt-3 perform statutory reasoning? In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, ICAIL '23, pp. 22–31, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701979. doi: 10.1145/3594536.3595163. URL https://doi.org/10.1145/3594536.3595163.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Ralph Allan Bradley and Milton Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3-4):324–345, 1952.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Fei Cai, Shangsong Liang, and Maarten De Rijke. Personalized document re-ranking based on bayesian probabilistic matrix factorization. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pp. 835–838, 2014.

Mark J Carman, Fabio Crestani, Morgan Harvey, and Mark Baillie. Towards query log based personalization using topic models. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pp. 1849–1852, 2010.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*, 2023.

Souradip Chakraborty, Soumya Suvra Ghosal, Ming Yin, Dinesh Manocha, Mengdi Wang, Amrit Singh Bedi, and Furong Huang. Transfer q star: Principled decoding for llm alignment, 2024.

Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*, 2024.

Zhiyuan Chang, Mingyang Li, Yi Liu, Junjie Wang, Qing Wang, and Yang Liu. Play guessing game with llm: Indirect jailbreak attack with implicit clues. *arXiv preprint arXiv:2402.09091*, 2024.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries. *arXiv preprint arXiv:2310.08419*, 2023.

Hongzhan Chen, Hehong Chen, Ming Yan, Wenshen Xu, Xing Gao, Weizhou Shen, Xiaojun Quan, Chenliang Li, Ji Zhang, Fei Huang, et al. Roleinteract: Evaluating the social interaction of role-playing agents. *arXiv preprint arXiv:2403.13679*, 2024a.

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

Jin Chen, Zheng Liu, Xu Huang, Chenwang Wu, Qi Liu, Gangwei Jiang, Yuanhao Pu, Yuxuan Lei, Xiaolong Chen, Xingmei Wang, et al. When large language models meet personalization: Perspectives of challenges and opportunities. *World Wide Web*, 27(4):42, 2024c.

Junyi Chen. A survey on large language models for personalized and explainable recommendations. *arXiv preprint arXiv:2311.12338*, 2023.

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. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Nuo Chen, Yan Wang, Haiyun Jiang, Deng Cai, Yuhan Li, Ziyang Chen, Longyue Wang, and Jia Li. Large language models meet harry potter: A dataset for aligning dialogue agents with characters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 8506–8520, 2023a.

Ruizhe Chen, Xiaotian Zhang, Meng Luo, Wenhao Chai, and Zuozhu Liu. Pad: Personalized alignment of llms at decoding-time. *arXiv preprint arXiv:2410.04070*, 2024d.

Yang Chen, Ethan Mendes, Sauvik Das, Wei Xu, and Alan Ritter. Can language models be instructed to protect personal information? *arXiv preprint arXiv:2310.02224*, 2023b.

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, et al. Scienceagentbench: Toward rigorous assessment of language agents for data-driven scientific discovery. *arXiv preprint arXiv:2410.05080*, 2024e.

Myra Cheng, Esin Durmus, and Dan Jurafsky. Marked personas: Using natural language prompts to measure stereotypes in language models. *arXiv preprint arXiv:2305.18189*, 2023a.

Myra Cheng, Tiziano Piccardi, and Diyi Yang. CoMPosT: Characterizing and evaluating caricature in LLM simulations. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10853–10875, Singapore, December 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.669. URL https://aclanthology.org/2023.emnlp-main.669.

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

Itsugun Cho, Dongyang Wang, Ryota Takahashi, and Hiroaki Saito. A personalized dialogue generator with implicit user persona detection. *arXiv preprint arXiv:2204.07372*, 2022.

Konstantina Christakopoulou, Alberto Lalama, Cj Adams, Iris Qu, Yifat Amir, Samer Chucri, Pierce Vollucci, Fabio Soldo, Dina Bseiso, Sarah Scodel, et al. Large language models for user interest journeys. *arXiv preprint arXiv:2305.15498*, 2023.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *arXiv preprint arXiv:2306.16092*, 2023.

Sunhao Dai, Ninglu Shao, Haiyuan Zhao, Weijie Yu, Zihua Si, Chen Xu, Zhongxiang Sun, Xiao Zhang, and Jun Xu. Uncovering chatgpt's capabilities in recommender systems. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pp. 1126–1132, 2023.

Zhenlong Dai, Chang Yao, WenKang Han, Yuanying Yuanying, Zhipeng Gao, and Jingyuan Chen. MPCoder: Multi-user personalized code generator with explicit and implicit style representation learning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3765–3780, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.207.

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

Yuhao Dan, Jie Zhou, Qin Chen, Junfeng Tian, and Liang He. P-tailor: Customizing personality traits for language models via mixture of specialized lora experts. *arXiv preprint arXiv:2406.12548*, 2024.

Sourish Dasgupta, Ankush Chander, Parth Borad, Isha Motiyani, and Tanmoy Chakraborty. Perseval: Assessing personalization in text summarizers. *arXiv preprint arXiv:2407.00453*, 2024.

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

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. Toxicity in chatgpt: Analyzing persona-assigned language models. *arXiv preprint arXiv:2304.05335*, 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Dario Di Palma. Retrieval-augmented recommender system: Enhancing recommender systems with large language models. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pp. 1369–1373, 2023.

Dario Di Palma, Giovanni Maria Biancofiore, Vito Walter Anelli, Fedelucio Narducci, Tommaso Di Noia, and Eugenio Di Sciascio. Evaluating chatgpt as a recommender system: A rigorous approach. *arXiv preprint arXiv:2309.03613*, 2023.

Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. Can ai language models replace human participants? *Trends in Cognitive Sciences*, 27(7):597–600, 2023.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. The second conversational intelligence challenge (convai2). In *The NeurIPS'18 Competition: From Machine Learning to Intelligent Conversations*, pp. 187–208. Springer, 2020.

Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. Longrope: Extending llm context window beyond 2 million tokens. *arXiv preprint arXiv:2402.13753*, 2024.

Sumanth Doddapaneni, Mohammed Safi Ur Rahman Khan, Sshubam Verma, and Mitesh M Khapra. Finding blind spots in evaluator llms with interpretable checklists. *arXiv preprint arXiv:2406.13439*, 2024.

Zhicheng Dou, Ruihua Song, and Ji-Rong Wen. A large-scale evaluation and analysis of personalized search strategies. In *Proceedings of the 16th international conference on World Wide Web*, pp. 581–590, 2007.

Xinya Du and Heng Ji. Retrieval-augmented generative question answering for event argument extraction. *arXiv preprint arXiv:2211.07067*, 2022.

Yingpeng Du, Di Luo, Rui Yan, Xiaopei Wang, Hongzhi Liu, Hengshu Zhu, Yang Song, and Jie Zhang. Enhancing job recommendation through llm-based generative adversarial networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 8363–8371, 2024.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. Glm: General language model pretraining with autoregressive blank infilling. *arXiv preprint arXiv:2103.10360*, 2021.

Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. Alpacafarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, 36, 2024.

Horatiu Dumitru, Marek Gibiec, Negar Hariri, Jane Cleland-Huang, Bamshad Mobasher, Carlos Castro-Herrera, and Mehdi Mirakhorli. On-demand feature recommendations derived from mining public product descriptions. In *Proceedings of the 33rd international conference on software engineering*, pp. 181–190, 2011.

Esin Durmus, Karina Nguyen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*, 2023.

Sabina Elkins, Ekaterina Kochmar, Iulian Serban, and Jackie CK Cheung. How useful are educational questions generated by large language models? In *International Conference on Artificial Intelligence in Education*, pp. 536–542. Springer, 2023.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. Understanding dataset difficulty with $\mathcal{V}$-usable information. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 5988–6008. PMLR, 17–23 Jul 2022.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, pp. 6491–6501, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704901. doi: 10.1145/3637528.3671470. URL https://doi.org/10.1145/3637528.3671470.

Zhiwei Fei, Songyang Zhang, Xiaoyu Shen, Dawei Zhu, Xiao Wang, Maosong Cao, Fengzhe Zhou, Yining Li, Wenwei Zhang, Dahua Lin, et al. Internlm-law: An open source chinese legal large language model. *arXiv preprint arXiv:2406.14887*, 2024.

Qizhang Feng, Siva Rajesh Kasa, Hyokun Yun, Choon Hui Teo, and Sravan Babu Bodapati. Exposing privacy gaps: Membership inference attack on preference data for llm alignment. *arXiv preprint arXiv:2407.06443*, 2024.

Iason Gabriel. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3):411–437, 2020.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics*, pp. 1–83, 07 2024. ISSN 0891-2017. doi: 10.1162/coli_a_00524. URL https://doi.org/10.1162/coli_a_00524.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.

Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. In *Proceedings of the 2018 world wide web conference*, pp. 913–922, 2018.

Suyu Ge, Chunting Zhou, Rui Hou, Madian Khabsa, Yi-Chia Wang, Qifan Wang, Jiawei Han, and Yuning Mao. Mart: Improving llm safety with multi-round automatic red-teaming. *arXiv preprint arXiv:2311.07689*, 2023.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*, 2020.

Somayeh Ghanbarzadeh, Yan Huang, Hamid Palangi, Radames Cruz Moreno, and Hamed Khanpour. Gender-tuning: Empowering fine-tuning for debiasing pre-trained language models. *arXiv preprint arXiv:2307.10522*, 2023.

Hannah Gonzalez, Jiening Li, Helen Jin, Jiaxuan Ren, Hongyu Zhang, Ayotomiwa Akinyele, Adrian Wang, Eleni Miltsakaki, Ryan Baker, and Chris Callison-Burch. 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)*, pp. 382–393, 2023.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, et al. Olmo: Accelerating the science of language models. *arXiv preprint arXiv:2402.00838*, 2024.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*, 2024.

Hui Guo. Soap: Live recommendations through social agents. In *Fifth DELOS Workshop on Filtering and Collaborative Filtering, Budapest*. Citeseer, 1997.

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

Muhammad Usman Hadi, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, et al. Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. *Authorea Preprints*, 2023.

Seungju Han, Beomsu Kim, Jin Yong Yoo, Seokjun Seo, Sangbum Kim, Enkhbayar Erdenee, and Buru Chang. Meet your favorite character: Open-domain chatbot mimicking fictional characters with only a few utterances. *arXiv preprint arXiv:2204.10825*, 2022.

Xudong Han, Timothy Baldwin, and Trevor Cohn. Balancing out bias: Achieving fairness through balanced training. *arXiv preprint arXiv:2109.08253*, 2021.

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

Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. The political ideology of conversational ai: Converging evidence on chatgpt's pro-environmental, left-libertarian orientation. *arXiv preprint arXiv:2301.01768*, 2023.

Morgan Harvey, Fabio Crestani, and Mark J Carman. Building user profiles from topic models for personalised search. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pp. 2309–2314, 2013.

Jerry Zhi-Yang He, Sashrika Pandey, Mariah L Schrum, and Anca Dragan. Cos: Enhancing personalization and mitigating bias with context steering. *arXiv preprint arXiv:2405.01768*, 2024a.

Zhicheng He, Weiwen Liu, Wei Guo, Jiarui Qin, Yingxue Zhang, Yaochen Hu, and Ruiming Tang. A survey on user behavior modeling in recommender systems. *arXiv preprint arXiv:2302.11087*, 2023.

Zhitao He, Pengfei Cao, Chenhao Wang, Zhuoran Jin, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. Simucourt: Building judicial decision-making agents with real-world judgement documents. *arXiv preprint arXiv:2403.02959*, 2024b.

Thomas F Heston and Charya Khun. Prompt engineering in medical education. *International Medical Education*, 2(3):198–205, 2023.

Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939*, 2015.

John J Horton. Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research, 2023.

Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. Large language models are zero-shot rankers for recommender systems. In *European Conference on Information Retrieval*, pp. 364–381. Springer, 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Zhengyu Hu, Linxin Song, Jieyu Zhang, Zheyuan Xiao, Jingang Wang, Zhenyu Chen, Jieyu Zhao, and Hui Xiong. Rethinking llm-based preference evaluation. *arXiv preprint arXiv:2407.01085*, 2024.

Hui Huang, Yingqi Qu, Jing Liu, Muyun Yang, and Tiejun Zhao. An empirical study of llm-as-a-judge for llm evaluation: Fine-tuned judge models are task-specific classifiers. *arXiv preprint arXiv:2403.02839*, 2024a.

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

Jen-tse Huang, Wenxuan Wang, Eric John Li, Man Ho Lam, Shujie Ren, Youliang Yuan, Wenxiang Jiao, Zhaopeng Tu, and Michael R Lyu. Who is chatgpt? benchmarking llms' psychological portrayal using psychobench. *arXiv preprint arXiv:2310.01386*, 2023b.

Qiushi Huang, Xubo Liu, Tom Ko, Bo Wu, Wenwu Wang, Yu Zhang, and Lilian Tang. Selective prompting tuning for personalized conversations with llms, 2024b. URL https://arxiv.org/abs/2406.18187.

Quzhe Huang, Mingxu Tao, Chen Zhang, Zhenwei An, Cong Jiang, Zhibin Chen, Zirui Wu, and Yansong Feng. Lawyer llama technical report. *arXiv preprint arXiv:2305.15062*, 2023c.

Stefan E Huber, Kristian Kiili, Steve Nebel, Richard M Ryan, Michael Sailer, and Manuel Ninaus. Leveraging the potential of large language models in education through playful and game-based learning. *Educational Psychology Review*, 36(1):25, 2024.

Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 874–880, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.74. URL https://aclanthology.org/2021.eacl-main.74.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*, 2021.

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

Sangwon Jang, Jaehyeong Jo, Kimin Lee, and Sung Ju Hwang. Identity decoupling for multi-subject personalization of text-to-image models. *arXiv preprint arXiv:2404.04243*, 2024.

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

Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *Advances in Neural Information Processing Systems*, 36, 2024a.

Jianchao Ji, Zelong Li, Shuyuan Xu, Wenyue Hua, Yingqiang Ge, Juntao Tan, and Yongfeng Zhang. Genrec: Large language model for generative recommendation. In *European Conference on Information Retrieval*, pp. 494–502. Springer, 2024b.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. Evaluating and inducing personality in pre-trained language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 10622–10643. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/21f7b745f73ce0d1f9bcea7f40b1388e-Paper-Conference.pdf.

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

Yucheng Jiang, Yijia Shao, Dekun Ma, Sina J Semnani, and Monica S Lam. Into the unknown unknowns: Engaged human learning through participation in language model agent conversations. *arXiv preprint arXiv:2408.15232*, 2024b.

Hongye Jin, Xiaotian Han, Jingfeng Yang, Zhimeng Jiang, Zirui Liu, Chia-Yuan Chang, Huiyuan Chen, and Xia Hu. Llm maybe longlm: Self-extend llm context window without tuning. *arXiv preprint arXiv:2401.01325*, 2024a.

Mingyu Jin, Qinkai Yu, Dong Shu, Chong Zhang, Lizhou Fan, Wenyue Hua, Suiyuan Zhu, Yanda Meng, Zhenting Wang, Mengnan Du, et al. Health-llm: Personalized retrieval-augmented disease prediction system. *arXiv preprint arXiv:2402.00746*, 2024b.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547, 2021. doi: 10.1109/TBDATA.2019.2921572.

Hideaki Joko, Shubham Chatterjee, Andrew Ramsay, Arjen P. de Vries, Jeff Dalton, and Faegheh Hasibi. Doing personal laps: Llm-augmented dialogue construction for personalized multi-session conversational search. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, pp. 796–806, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704314. doi: 10.1145/3626772.3657815. URL https://doi.org/10.1145/3626772.3657815.

Ishika Joshi, Ritvik Budhiraja, Pranav Deepak Tanna, Lovenya Jain, Mihika Deshpande, Arjun Srivastava, Srinivas Rallapalli, Harshal D Akolekar, Jagat Sesh Challa, and Dhruv Kumar. From" let's google" to" let's chatgpt": Student and instructor perspectives on the influence of llms on undergraduate engineering education. *arXiv preprint arXiv:2309.10694*, 2023.

Wang-Cheng Kang, Jianmo Ni, Nikhil Mehta, Maheswaran Sathiamoorthy, Lichan Hong, Ed Chi, and Derek Zhiyuan Cheng. Do llms understand user preferences? evaluating llms on user rating prediction. *arXiv preprint arXiv:2305.06474*, 2023.

Anjali Kantharuban, Jeremiah Milbauer, Emma Strubell, and Graham Neubig. Stereotype or personalization? user identity biases chatbot recommendations. *arXiv preprint arXiv:2410.05613*, 2024.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.550. URL https://aclanthology.org/2020.emnlp-main.550.

Saketh Reddy Karra, Son The Nguyen, and Theja Tulabandhula. Estimating the personality of white-box language models. *arXiv preprint arXiv:2204.12000*, 2022.

Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, Stephan Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, Matthias Stadler, Jochen Weller, Jochen Kuhn, and Gjergji Kasneci. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274, 2023. ISSN 1041-6080. doi: https://doi.org/10.1016/j.lindif.2023.102274. URL https://www.sciencedirect.com/science/article/pii/S1041608023000195.

Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. A survey of reinforcement learning from human feedback. *arXiv preprint arXiv:2312.14925*, 2023.

Jaehyung Kim and Yiming Yang. Few-shot personalization of llms with mis-aligned responses. *arXiv preprint arXiv:2406.18678*, 2024.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Prometheus 2: An open source language model specialized in evaluating other language models. *arXiv preprint arXiv:2405.01535*, 2024.

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, et al. The prism alignment project: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *arXiv preprint arXiv:2404.16019*, 2024.

Grgur Kovač, Masataka Sawayama, Rémy Portelas, Cédric Colas, Peter Ford Dominey, and Pierre-Yves Oudeyer. Large language models as superpositions of cultural perspectives. *arXiv preprint arXiv:2307.07870*, 2023.

Weirui Kuang, Bingchen Qian, Zitao Li, Daoyuan Chen, Dawei Gao, Xuchen Pan, Yuexiang Xie, Yaliang Li, Bolin Ding, and Jingren Zhou. Federatedscope-llm: A comprehensive package for fine-tuning large language models in federated learning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery*

and Data Mining, KDD '24, pp. 5260–5271, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704901. doi: 10.1145/3637528.3671573. URL https://doi.org/10.1145/3637528.3671573.

Viraj Kulkarni, Milind Kulkarni, and Aniruddha Pant. Survey of personalization techniques for federated learning. In *2020 fourth world conference on smart trends in systems, security and sustainability (WorldS4)*, pp. 794–797. IEEE, 2020.

Ishita Kumar, Snigdha Viswanathan, Sushrita Yerra, Ryan A. Rossi Alireza Salemi, Franck Dernoncourt, Hanieh Deilamsalehy, Xiang Chen, Ruiyi Zhang, Shubham Agarwal, Nedim Lipka, and Hamed Zamani. Longlamp: A benchmark for personalized long-form text generation. *arXiv preprint*, 2024.

Jinqi Lai, Wensheng Gan, Jiayang Wu, Zhenlian Qi, and Philip S Yu. Large language models in law: A survey. *arXiv preprint arXiv:2312.03718*, 2023.

Kausik Lakkaraju, Sai Krishna Revanth Vuruma, Vishal Pallagani, Bharath Muppasani, and Biplav Srivastava. Can llms be good financial advisors?: An initial study in personal decision making for optimized outcomes. *arXiv preprint arXiv:2307.07422*, 2023.

Xuan Nhat Lam, Thuc Vu, Trong Duc Le, and Anh Duc Duong. Addressing cold-start problem in recommendation systems. In *Proceedings of the 2nd International Conference on Ubiquitous Information Management and Communication*, ICUIMC '08, pp. 208–211, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781595939937. doi: 10.1145/1352793.1352837. URL https://doi.org/10.1145/1352793.1352837.

Tomo Lazovich. Filter bubbles and affective polarization in user-personalized large language model outputs. In Javier Antorán, Arno Blaas, Kelly Buchanan, Fan Feng, Vincent Fortuin, Sahra Ghalebikesabi, Andreas Kriegler, Ian Mason, David Rohde, Francisco J. R. Ruiz, Tobias Uelwer, Yubin Xie, and Rui Yang (eds.), *Proceedings on "I Can't Believe It's Not Better: Failure Modes in the Age of Foundation Models" at NeurIPS 2023 Workshops*, volume 239 of *Proceedings of Machine Learning Research*, pp. 29–37. PMLR, 16 Dec 2023. URL https://proceedings.mlr.press/v239/lazovich23a.html.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Ren Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, et al. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback. In *Forty-first International Conference on Machine Learning*.

Hoyeop Lee, Jinbae Im, Seongwon Jang, Hyunsouk Cho, and Sehee Chung. Melu: Meta-learned user preference estimator for cold-start recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1073–1082, 2019.

Seongyun Lee, Sue Hyun Park, Seungone Kim, and Minjoon Seo. Aligning to thousands of preferences via system message generalization. *arXiv preprint arXiv:2405.17977*, 2024.

Joanne Leong, Pat Pataranutaporn, Valdemar Danry, Florian Perteneder, Yaoli Mao, and Pattie Maes. Putting things into context: Generative ai-enabled context personalization for vocabulary learning improves learning motivation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703300. doi: 10.1145/3613904.3642393. URL https://doi.org/10.1145/3613904.3642393.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi Mi, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, et al. Chatharuhi: Reviving anime character in reality via large language model. *arXiv preprint arXiv:2308.09597*, 2023a.

Cheng Li, Mingyang Zhang, Qiaozhu Mei, Yaqing Wang, Spurthi Amba Hombaiah, Yi Liang, and Michael Bendersky. Teach llms to personalize–an approach inspired by writing education. *arXiv preprint arXiv:2308.07968*, 2023b.

Cheng Li, Mingyang Zhang, Qiaozhu Mei, Weize Kong, and Michael Bendersky. Learning to rewrite prompts for personalized text generation. In *Proceedings of the ACM on Web Conference 2024*, WWW '24. ACM, May 2024a. doi: 10.1145/3589334.3645408. URL http://dx.doi.org/10.1145/3589334.3645408.

Cheng Li, Mingyang Zhang, Qiaozhu Mei, Weize Kong, and Michael Bendersky. Learning to rewrite prompts for personalized text generation. In *Proceedings of the ACM on Web Conference 2024*, pp. 3367–3378, 2024b.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, et al. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *arXiv preprint arXiv:2411.16594*, 2024c.

Jiarui Li, Ye Yuan, and Zehua Zhang. Enhancing llm factual accuracy with rag to counter hallucinations: A case study on domain-specific queries in private knowledge-bases. *arXiv preprint arXiv:2403.10446*, 2024d.

Jingjing Li, Mengmeng Jing, Ke Lu, Lei Zhu, Yang Yang, and Zi Huang. From zero-shot learning to cold-start recommendation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 4189–4196, 2019.

Jinyang Li, Nan Huo, Yan Gao, Jiayi Shi, Yingxiu Zhao, Ge Qu, Yurong Wu, Chenhao Ma, Jian-Guang Lou, and Reynold Cheng. Tapilot-crossing: Benchmarking and evolving llms towards interactive data analysis agents. *arXiv preprint arXiv:2403.05307*, 2024e.

Lei Li, Yongfeng Zhang, and Li Chen. Generate neural template explanations for recommendation. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, pp. 755–764, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450368599. doi: 10.1145/3340531.3411992. URL https://doi.org/10.1145/3340531.3411992.

Lei Li, Yongfeng Zhang, and Li Chen. Prompt distillation for efficient llm-based recommendation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, pp. 1348–1357, New York, NY, USA, 2023c. Association for Computing Machinery. ISBN 9798400701245. doi: 10.1145/3583780.3615017. URL https://doi.org/10.1145/3583780.3615017.

Minzhi Li, Taiwei Shi, Caleb Ziems, Min-Yen Kan, Nancy Chen, Zhengyuan Liu, and Diyi Yang. CoAnnotating: Uncertainty-guided work allocation between human and large language models for data annotation. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 1487–1505, Singapore, December 2023d. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.92. URL https://aclanthology.org/2023.emnlp-main.92.

Pan Li and Alexander Tuzhilin. Towards controllable and personalized review generation. *arXiv preprint arXiv:1910.03506*, 2019.

Sheng Li and Handong Zhao. A survey on representation learning for user modeling. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pp. 4997–5003, 2021.

Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *International conference on machine learning*, pp. 6357–6368. PMLR, 2021.

Wei Li, Xue Xu, Jiachen Liu, and Xinyan Xiao. Unimo-g: Unified image generation through multimodal conditional diffusion. *arXiv preprint arXiv:2401.13388*, 2024f.

Xinyi Li, Yongfeng Zhang, and Edward C Malthouse. A preliminary study of chatgpt on news recommendation: Personalization, provider fairness, fake news. *arXiv preprint arXiv:2306.10702*, 2023e.

Xinyu Li, Zachary C Lipton, and Liu Leqi. Personalized language modeling from personalized human feedback. *arXiv preprint arXiv:2402.05133*, 2024g.

Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. A survey on fairness in large language models. *arXiv preprint arXiv:2308.10149*, 2023f.

Yuanchun Li, Hao Wen, Weijun Wang, Xiangyu Li, Yizhen Yuan, Guohong Liu, Jiacheng Liu, Wenxing Xu, Xiang Wang, Yi Sun, et al. Personal llm agents: Insights and survey about the capability, efficiency and security. *arXiv preprint arXiv:2401.05459*, 2024h.

Zongjie Li, Chaozheng Wang, Pingchuan Ma, Daoyuan Wu, Shuai Wang, Cuiyun Gao, and Yang Liu. Split and merge: Aligning position biases in LLM-based evaluators. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 11084–11108, Miami, Florida, USA, November 2024i. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.621. URL https://aclanthology.org/2024.emnlp-main.621/.

Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. The unlocking spell on base LLMs: Rethinking alignment via in-context learning. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=wxJ0eXwwda.

Bin Lin, Tao Peng, Chen Zhang, Minmin Sun, Lanbo Li, Hanyu Zhao, Wencong Xiao, Qi Xu, Xiafei Qiu, Shen Li, et al. Infinite-llm: Efficient llm service for long context with distattention and distributed kvcache. *arXiv preprint arXiv:2401.02669*, 2024b.

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

Chin-Yew Lin and Eduard Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 human language technology conference of the North American chapter of the association for computational linguistics*, pp. 150–157, 2003.

Chin-Yew Lin and Franz Josef Och. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd annual meeting of the association for computational linguistics (ACL-04)*, pp. 605–612, 2004.

Jiongnan Liu, Yutao Zhu, Shuting Wang, Xiaochi Wei, Erxue Min, Yu Lu, Shuaiqiang Wang, Dawei Yin, and Zhicheng Dou. Llms+ persona-plug= personalized llms. *arXiv preprint arXiv:2409.11901*, 2024a.

Junling Liu, Chao Liu, Peilin Zhou, Renjie Lv, Kang Zhou, and Yan Zhang. Is chatgpt a good recommender? a preliminary study. *arXiv preprint arXiv:2304.10149*, 2023a.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024b. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024c.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9), jan 2023b. ISSN 0360-0300. doi: 10.1145/3560815. URL https://doi.org/10.1145/3560815.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023c.

Qijiong Liu, Nuo Chen, Tetsuya Sakai, and Xiao-Ming Wu. Once: Boosting content-based recommendation with both open-and closed-source large language models. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pp. 452–461, 2024d.

Ruibo Liu. Aligning language models with the human world. (276), 2024. URL https://digitalcommons.dartmouth.edu/dissertations/276.

Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, Lili Wang, and Soroush Vosoughi. Mitigating political bias in language models through reinforced calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 14857–14866, 2021.

Xiao-Yang Liu, Guoxuan Wang, and Daochen Zha. Fingpt: Democratizing internet-scale data for financial large language models. *arXiv preprint arXiv:2307.10485*, 2023d.

Xin Liu, Daniel McDuff, Geza Kovacs, Isaac Galatzer-Levy, Jacob Sunshine, Jiening Zhan, Ming-Zher Poh, Shun Liao, Paolo Di Achille, and Shwetak Patel. Large language models are few-shot health learners. *arXiv preprint arXiv:2305.15525*, 2023e.

Yi Liu, Gelei Deng, Yuekang Li, Kailong Wang, Zihao Wang, Xiaofeng Wang, Tianwei Zhang, Yepang Liu, Haoyu Wang, Yan Zheng, et al. Prompt injection attack against llm-integrated applications. *arXiv preprint arXiv:2306.05499*, 2023f.

Yi Liu, Gelei Deng, Zhengzi Xu, Yuekang Li, Yaowen Zheng, Ying Zhang, Lida Zhao, Tianwei Zhang, Kailong Wang, and Yang Liu. Jailbreaking chatgpt via prompt engineering: An empirical study. *arXiv preprint arXiv:2305.13860*, 2023g.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Yuqing Liu, Yu Wang, Lichao Sun, and Philip S Yu. Rec-gpt4v: Multimodal recommendation with large vision-language models. *arXiv preprint arXiv:2402.08670*, 2024e.

Zhengliang Liu, Zihao Wu, Mengxuan Hu, Bokai Zhao, Lin Zhao, Tianyi Zhang, Haixing Dai, Xianyan Chen, Ye Shen, Sheng Li, et al. Pharmacygpt: The ai pharmacist. *arXiv preprint arXiv:2307.10432*, 2023h.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning*, pp. 22631–22648. PMLR, 2023.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. Gender bias in neural natural language processing. *Logic, language, and security: essays dedicated to Andre Scedrov on the occasion of his 65th birthday*, pp. 189–202, 2020.

Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. Analyzing leakage of personally identifiable information in language models. In *2023 IEEE Symposium on Security and Privacy (SP)*, pp. 346–363. IEEE, 2023.

Hanjia Lyu and Jiebo Luo. Understanding political polarization via jointly modeling users, connections and multimodal contents on heterogeneous graphs. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 4072–4082, 2022.

Hanjia Lyu, Song Jiang, Hanqing Zeng, Yinglong Xia, and Jiebo Luo. LLM-Rec: Personalized recommendation via prompting large language models. *arXiv preprint arXiv:2307.15780*, 2023.

Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–12, 2024.

Bodhisattwa Prasad Majumder, Shuyang Li, Jianmo Ni, and Julian McAuley. Generating personalized recipes from historical user preferences. *arXiv preprint arXiv:1909.00105*, 2019.

Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. Training millions of personalized dialogue agents. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2775–2779, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1298. URL https://aclanthology.org/D18-1298.

Walaa Medhat, Ahmed Hassan, and Hoda Korashy. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113, 2014.

Bertalan Meskó. Prompt engineering as an important emerging skill for medical professionals: tutorial. *Journal of medical Internet research*, 25:e50638, 2023.

Microsoft. Reinventing search with a new ai-powered bing and edge, your copilot for the web, 2023. URL https://news.microsoft.com/the-new-bing/. Accessed: 2024-07-29.

Microsoft. Meet microsoft copilot. https://www.microsoft.com/en-us/microsoft-copilot/meet-copilot, 2024. Accessed: 2024-09-17.

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

Alan L Montgomery, Shibo Li, Kannan Srinivasan, and John C Liechty. Modeling online browsing and path analysis using clickstream data. *Marketing science*, 23(4):579–595, 2004.

Sheshera Mysore, Zhuoran Lu, Mengting Wan, Longqi Yang, Steve Menezes, Tina Baghaee, Emmanuel Barajas Gonzalez, Jennifer Neville, and Tara Safavi. Pearl: Personalizing large language model writing assistants with generation-calibrated retrievers. *arXiv preprint arXiv:2311.09180*, 2023a.

Sheshera Mysore, Andrew McCallum, and Hamed Zamani. Large language model augmented narrative driven recommendations. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pp. 777–783, 2023b.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

John J Nay. Large language models as fiduciaries: a case study toward robustly communicating with artificial intelligence through legal standards. *arXiv preprint arXiv:2301.10095*, 2023.

Andrew Y Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Icml*, volume 99, pp. 278–287, 1999.

Ha-Thanh Nguyen. A brief report on lawgpt 1.0: A virtual legal assistant based on gpt-3. *arXiv preprint arXiv:2302.05729*, 2023.

Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 188–197, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1018. URL https://aclanthology.org/D19-1018.

Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith B Hall, Daniel Cer, and Yinfei Yang. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. *arXiv preprint arXiv:2108.08877*, 2021.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang, and Yinfei Yang. Large dual encoders are generalizable retrievers. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 9844–9855, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.669. URL https://aclanthology.org/2022.emnlp-main.669.

Liang-bo Ning, Shijie Wang, Wenqi Fan, Qing Li, Xin Xu, Hao Chen, and Feiran Huang. Cheatagent: Attacking llm-empowered recommender systems via llm agent. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, pp. 2284–2295, New York, NY, USA, 2024a. Association for Computing Machinery. ISBN 9798400704901. doi: 10.1145/3637528.3671837. URL https://doi.org/10.1145/3637528.3671837.

Lin Ning, Luyang Liu, Jiaxing Wu, Neo Wu, Devora Berlowitz, Sushant Prakash, Bradley Green, Shawn O'Banion, and Jun Xie. User-llm: Efficient llm contextualization with user embeddings. *arXiv preprint arXiv:2402.13598*, 2024b.

Vít Novotnỳ and Michal Stefánik. Combining sparse and dense information retrieval. In *CLEF (Working Notes)*, pp. 104–118, 2022.

Peter Ochieng. Are large language models fit for guided reading? *arXiv preprint arXiv:2305.10645*, 2023.

Anastasia Olga, Akash Saini, Gabriela Zapata, Duane Searsmith, Bill Cope, Mary Kalantzis, Vania Castro, Theodora Kourkoulou, John Jones, Rodrigo Abrantes da Silva, et al. Generative ai: Implications and applications for education. *arXiv preprint arXiv:2305.07605*, 2023.

Damilola Omitaomu, Shabnam Tafreshi, Tingting Liu, Sven Buechel, Chris Callison-Burch, Johannes Eichstaedt, Lyle Ungar, and João Sedoc. Empathic conversations: A multi-level dataset of contextualized conversations. *arXiv preprint arXiv:2205.12698*, 2022.

OpenAI. Searchgpt is a prototype of new ai search features, 2024. URL https://openai.com/index/searchgpt-prototype/. Accessed: 2024-07-29.

Matthias Orlikowski, Paul Röttger, Philipp Cimiano, and Dirk Hovy. The ecological fallacy in annotation: Modelling human label variation goes beyond sociodemographics. *arXiv preprint arXiv:2306.11559*, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

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

Eli Pariser. *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin, 2011.

Chanwoo Park, Mingyang Liu, Kaiqing Zhang, and Asuman Ozdaglar. Principled rlhf from heterogeneous feedback via personalization and preference aggregation. *arXiv preprint arXiv:2405.00254*, 2024a.

Minju Park, Sojung Kim, Seunghyun Lee, Soonwoo Kwon, and Kyuseok Kim. Empowering personalized learning through a conversation-based tutoring system with student modeling. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI EA '24, New York, NY, USA, 2024b. Association for Computing Machinery. ISBN 9798400703317. doi: 10.1145/3613905.3651122. URL https://doi.org/10.1145/3613905.3651122.

Seung-Taek Park and Wei Chu. Pairwise preference regression for cold-start recommendation. In *Proceedings of the third ACM conference on Recommender systems*, pp. 21–28, 2009.

Luis G Perez, Manuel Barranco, and Luis Martinez. Building user profiles for recommender systems from incomplete preference relations. In *2007 IEEE International Fuzzy Systems Conference*, pp. 1–6. IEEE, 2007.

Tung Phung, Victor-Alexandru Pădurean, José Cambronero, Sumit Gulwani, Tobias Kohn, Rupak Majumdar, Adish Singla, and Gustavo Soares. Generative ai for programming education: Benchmarking chatgpt, gpt-4, and human tutors. In *Proceedings of the 2023 ACM Conference on International Computing Education Research-Volume 2*, pp. 41–42, 2023.

Robin Lewis Plackett. The analysis of permutations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 24(2):193–202, 1975.

Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. Personalizing reinforcement learning from human feedback with variational preference learning. *arXiv preprint arXiv:2408.10075*, 2024.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. Is ChatGPT a general-purpose natural language processing task solver? In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 1339–1384, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.85. URL https://aclanthology.org/2023.emnlp-main.85.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Vyas Raina, Adian Liusie, and Mark Gales. Is llm-as-a-judge robust? investigating universal adversarial attacks on zero-shot llm assessment. *arXiv preprint arXiv:2402.14016*, 2024.

Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan Hulikal Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Tran, Jonah Samost, Maciej Kula, Ed Chi, and Maheswaran Sathiamoorthy. Recommender systems with generative retrieval. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 10299–10315. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/20dcab0f14046a5c6b02b61da9f13229-Paper-Conference.pdf.

Abhinav Rao, Aditi Khandelwal, Kumar Tanmay, Utkarsh Agarwal, and Monojit Choudhury. Ethical reasoning over moral alignment: A case and framework for in-context ethical policies in llms. *arXiv preprint arXiv:2310.07251*, 2023.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410.

Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. Grouplens: An open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM conference on Computer supported cooperative work*, pp. 175–186, 1994.

Chris Richardson, Yao Zhang, Kellen Gillespie, Sudipta Kar, Arshdeep Singh, Zeynab Raeesy, Omar Zia Khan, and Abhinav Sethy. Integrating summarization and retrieval for enhanced personalization via large language models. *arXiv preprint arXiv:2310.20081*, 2023.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. Okapi at trec-3. *Nist Special Publication Sp*, 109:109, 1995.

Peter E Rossi, Robert E McCulloch, and Greg M Allenby. The value of purchase history data in target marketing. *Marketing Science*, 15(4):321–340, 1996.

Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, et al. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*, 2023.

Michael J Ryan, William Held, and Diyi Yang. Unintended impacts of llm alignment on global representation. *arXiv preprint arXiv:2402.15018*, 2024.

Mustafa Safdari, Greg Serapio-García, Clément Crepy, Stephen Fitz, Peter Romero, Luning Sun, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. Personality traits in large language models. *arXiv preprint arXiv:2307.00184*, 2023.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10776–10787, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.722. URL https://aclanthology.org/2023.findings-emnlp.722/.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. Lamp: When large language models meet personalization. *arXiv preprint arXiv:2304.11406*, 2023.

Alireza Salemi, Surya Kallumadi, and Hamed Zamani. Optimization methods for personalizing large language models through retrieval augmentation, 2024.

Pierangela Samarati and Latanya Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. 1998.

Scott Sanner, Krisztian Balog, Filip Radlinski, Ben Wedin, and Lucas Dixon. Large language models are competitive near cold-start recommenders for language-and item-based preferences. In *Proceedings of the 17th ACM conference on recommender systems*, pp. 890–896, 2023.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pp. 29971–30004. PMLR, 2023.

J Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen. Collaborative filtering recommender systems. In *The adaptive web: methods and strategies of web personalization*, pp. 291–324. Springer, 2007.

Andrew I. Schein, Alexandrin Popescul, Lyle H. Ungar, and David M. Pennock. Methods and metrics for cold-start recommendations. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '02, pp. 253–260, New York, NY, USA, 2002. Association for Computing Machinery. ISBN 1581135610. doi: 10.1145/564376.564421. URL https://doi.org/10.1145/564376.564421.

Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, HyoJung Han, Sevien Schulhoff, et al. The prompt report: A systematic survey of prompting techniques. *arXiv preprint arXiv:2406.06608*, 2024.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 1889–1897, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Omar Shaikh, Michelle Lam, Joey Hejna, Yijia Shao, Michael Bernstein, and Diyi Yang. Show, don't tell: Aligning language models with demonstrated feedback. *arXiv preprint arXiv:2406.00888*, 2024.

Shreya Shankar, JD Zamfirescu-Pereira, Björn Hartmann, Aditya G Parameswaran, and Ian Arawjo. Who validates the validators? aligning llm-assisted evaluation of llm outputs with human preferences. *arXiv preprint arXiv:2404.12272*, 2024.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. Character-LLM: A trainable agent for role-playing. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13153–13187, Singapore, December 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.emnlp-main.814.

Nikhil Sharma, Q. Vera Liao, and Ziang Xiao. Generative echo chamber? effect of llm-powered search systems on diverse information seeking. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703300. doi: 10.1145/3613904.3642459. URL https://doi.org/10.1145/3613904.3642459.

Prabin Sharma, Kisan Thapa, Dikshya Thapa, Prastab Dhakal, Mala Deep Upadhaya, Santosh Adhikari, and Salik Ram Khanal. Performance of chatgpt on usmle: Unlocking the potential of large language models for ai-assisted medical education. *arXiv preprint arXiv:2307.00112*, 2023.

Tianhao Shen, Sun Li, and Deyi Xiong. Roleeval: A bilingual role evaluation benchmark for large language models. *arXiv preprint arXiv:2312.16132*, 2023.

Xiaoteng Shen, Rui Zhang, Xiaoyan Zhao, Jieming Zhu, and Xi Xiao. Pmg: Personalized multimodal generation with large language models. In *Proceedings of the ACM on Web Conference 2024*, pp. 3833–3843, 2024.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pp. 31210–31227. PMLR, 2023.

Jiawen Shi, Zenghui Yuan, Yinuo Liu, Yue Huang, Pan Zhou, Lichao Sun, and Neil Zhenqiang Gong. Optimization-based prompt injection attack to llm-as-a-judge. *arXiv preprint arXiv:2403.17710*, 2024a.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Raya Horesh, Rogério Abreu de Paula, Diyi Yang, et al. Culturebank: An online community-driven knowledge base towards culturally aware language technologies. *arXiv preprint arXiv:2404.15238*, 2024b.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*, 2021.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. *arXiv preprint arXiv:2409.04109*, 2024.

Ahu Sieg, Bamshad Mobasher, and Robin Burke. Web search personalization with ontological user profiles. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pp. 525–534, 2007.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33:16857–16867, 2020.

Kunpeng Song, Yizhe Zhu, Bingchen Liu, Qing Yan, Ahmed Elgammal, and Xiao Yang. Moma: Multimodal llm adapter for fast personalized image generation. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 117–132, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-73661-2.

Yang Song, Hongning Wang, and Xiaodong He. Adapting deep ranknet for personalized search. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pp. 83–92, 2014.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. A roadmap to pluralistic alignment, 2024.

Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21, 1972.

Sofia Eleni Spatharioti, David M Rothschild, Daniel G Goldstein, and Jake M Hofman. Comparing traditional and llm-based search for consumer choice: A randomized experiment. *arXiv preprint arXiv:2307.03744*, 2023.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. Beyond memorization: Violating privacy via inference with large language models. *arXiv preprint arXiv:2310.07298*, 2023.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback. In *NeurIPS*, 2020a.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020b.

Chenkai Sun, Ke Yang, Revanth Gangi Reddy, Yi R Fung, Hou Pong Chan, ChengXiang Zhai, and Heng Ji. Persona-db: Efficient large language model personalization for response prediction with collaborative data refinement. *arXiv preprint arXiv:2402.11060*, 2024.

Richard Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 3:9–44, 1988.

Richard Sutton and Andrew Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.

Richard Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems 12*, pp. 1057–1063, 2000.

Zhaoxuan Tan and Meng Jiang. User modeling in the era of large language models: Current research and future directions, 2023.

Zhaoxuan Tan, Zheyuan Liu, and Meng Jiang. Personalized pieces: Efficient personalized large language models through collaborative efforts. *arXiv preprint arXiv:2406.10471*, 2024a.

Zhaoxuan Tan, Qingkai Zeng, Yijun Tian, Zheyuan Liu, Bing Yin, and Meng Jiang. Democratizing large language models via personalized parameter-efficient fine-tuning. *arXiv preprint arXiv:2402.04401*, 2024b.

Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. Does synthetic data generation of llms help clinical text mining?, 2023. URL https://arxiv.org/abs/2303.04360.

Jessica Taylor, Eliezer Yudkowsky, Patrick LaVictoire, and Andrew Critch. Alignment for advanced machine learning systems. *Ethics of artificial intelligence*, pp. 342–382, 2016.

Jaime Teevan, Daniel J Liebling, and Gayathri Ravichandran Geetha. Understanding and predicting personal navigation. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pp. 85–94, 2011.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *arXiv preprint arXiv:2305.14975*, 2023.

Yuxuan Tong, Xiwen Zhang, Rui Wang, Ruidong Wu, and Junxian He. Dart-math: Difficulty-aware rejection tuning for mathematical problem-solving. *arXiv preprint arXiv:2407.13690*, 2024.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Dietrich Trautmann, Alina Petrova, and Frank Schilder. Legal prompt engineering for multilingual legal judgement prediction. *arXiv preprint arXiv:2212.02199*, 2022.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Yu-Ching Hsu, Jia-Yin Foo, Chao-Wei Huang, and Yun-Nung Chen. Two tales of persona in llms: A survey of role-playing and personalization. *arXiv preprint arXiv:2406.01171*, 2024.

Xinming Tu, James Zou, Weijie J Su, and Linjun Zhang. What should data science education do with large language models. *arXiv preprint arXiv:2307.02792*, 3, 2023.

Rahul Vansh, Darsh Rank, Sourish Dasgupta, and Tanmoy Chakraborty. Accuracy is not enough: Evaluating personalization in summarizers. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 2582–2595, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.169. URL https://aclanthology.org/2023.findings-emnlp.169.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Anvesh Rao Vijjini, Somnath Basu Roy Chowdhury, and Snigdha Chaturvedi. Exploring safety-utility trade-offs in personalized language models. *arXiv preprint arXiv:2406.11107*, 2024.

Thanh Vu, Alistair Willis, Son N Tran, and Dawei Song. Temporal latent topic user profiles for search personalisation. In *Advances in Information Retrieval: 37th European Conference on IR Research, ECIR 2015, Vienna, Austria, March 29-April 2, 2015. Proceedings 37*, pp. 605–616. Springer, 2015.

Thanh Tien Vu, Dawei Song, Alistair Willis, Son Ngoc Tran, and Jingfei Li. Improving search personalisation with dynamic group formation. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pp. 951–954, 2014.

Yixin Wan, Jieyu Zhao, Nanyun Peng, Kai-Wei Chang, and Aman Chadha. Are personalized stochastic parrots more dangerous? evaluating persona biases in dialogue systems. *arXiv preprint arXiv:2310.05280*, 2023.

Angelina Wang, Jamie Morgenstern, and John P Dickerson. Large language models cannot replace human participants because they cannot portray identity groups. *arXiv preprint arXiv:2402.01908*, 2024a.

Jindong Wang, Xixu Hu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Haojun Huang, Wei Ye, Xiubo Geng, et al. On the robustness of chatgpt: An adversarial and out-of-distribution perspective. *arXiv preprint arXiv:2302.12095*, 2023a.

Lei Wang and Ee-Peng Lim. Zero-shot next-item recommendation using large pretrained language models. *arXiv preprint arXiv:2304.03153*, 2023.

Rose E Wang and Dorottya Demszky. Is chatgpt a good teacher coach? measuring zero-shot performance for scoring and providing actionable insights on classroom instruction. *arXiv preprint arXiv:2306.03090*, 2023.

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

Wenjie Wang, Xinyu Lin, Fuli Feng, Xiangnan He, and Tat-Seng Chua. Generative recommendation: Towards next-generation recommender paradigm. *arXiv preprint arXiv:2304.03516*, 2023b.

X Wang, Siming Fu, Qihan Huang, Wanggui He, and Hao Jiang. Ms-diffusion: Multi-subject zero-shot image personalization with layout guidance. *arXiv preprint arXiv:2406.07209*, 2024c.

Xingyao Wang, Yangyi Chen, Lifan Yuan, Yizhe Zhang, Yunzhu Li, Hao Peng, and Heng Ji. Executable code actions elicit better llm agents. *arXiv preprint arXiv:2402.01030*, 2024d.

Xingyao Wang, Boxuan Li, Yufan Song, Frank F Xu, Xiangru Tang, Mingchen Zhuge, Jiayi Pan, Yueqi Song, Bowen Li, Jaskirat Singh, et al. Opendevin: An open platform for ai software developers as generalist agents. *arXiv preprint arXiv:2407.16741*, 2024e.

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. Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews, 2024f.

Yu Wang, Nedim Lipka, Ryan A Rossi, Alexa Siu, Ruiyi Zhang, and Tyler Derr. Knowledge graph prompting for multi-document question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19206–19214, 2024g.

Yuqing Wang, Yun Zhao, and Linda Petzold. Are large language models ready for healthcare? a comparative study on clinical language understanding. In *Machine Learning for Healthcare Conference*, pp. 804–823. PMLR, 2023c.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Man Zhang, et al. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv:2310.00746*, 2023d.

Zichao Wang, Jakob Valdez, Debshila Basu Mallick, and Richard G. Baraniuk. *Towards Human-Like Educational Question Generation with Large Language Models*, pp. 153–166. Springer International Publishing, 2022. ISBN 9783031116445. doi: 10.1007/978-3-031-11644-5_13. URL http://dx.doi.org/10.1007/978-3-031-11644-5_13.

Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7):5731–5780, 2022.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? *Advances in Neural Information Processing Systems*, 36, 2024a.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022a.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022b.

Jian Wei, Jianhua He, Kai Chen, Yi Zhou, and Zuoyin Tang. Collaborative filtering and deep learning based recommendation system for cold start items. *Expert Systems with Applications*, 69:29–39, 2017.

Tianxin Wei, Bowen Jin, Ruirui Li, Hansi Zeng, Zhengyang Wang, Jianhui Sun, Qingyu Yin, Hanqing Lu, Suhang Wang, Jingrui He, et al. Towards unified multi-modal personalization: Large vision-language models for generative recommendation and beyond. *arXiv preprint arXiv:2403.10667*, 2024b.

Iain Weissburg, Sathvika Anand, Sharon Levy, and Haewon Jeong. Llms are biased teachers: Evaluating llm bias in personalized education. *arXiv preprint arXiv:2410.14012*, 2024.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*, 2023.

Ryen W White, Wei Chu, Ahmed Hassan, Xiaodong He, Yang Song, and Hongning Wang. Enhancing personalized search by mining and modeling task behavior. In *Proceedings of the 22nd international conference on World Wide Web*, pp. 1411–1420, 2013.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8:229–256, 1992.

Stanisław Woźniak, Bartłomiej Koptyra, Arkadiusz Janz, Przemysław Kazienko, and Jan Kocoń. Personalized large language models. *arXiv preprint arXiv:2402.09269*, 2024.

Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, et al. Mind: A large-scale dataset for news recommendation. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 3597–3606, 2020.

Junda Wu, Hanjia Lyu, Yu Xia, Zhehao Zhang, Joe Barrow, Ishita Kumar, Mehrnoosh Mirtaheri, Hongjie Chen, Ryan A Rossi, Franck Dernoncourt, et al. Personalized multimodal large language models: A survey. *arXiv preprint arXiv:2412.02142*, 2024a.

Likang Wu, Zhi Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, et al. A survey on large language models for recommendation. *arXiv preprint arXiv:2305.19860*, 2023a.

Likang Wu, Zhaopeng Qiu, Zhi Zheng, Hengshu Zhu, and Enhong Chen. Exploring large language model for graph data understanding in online job recommendations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 9178–9186, 2024b.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*, 2023b.

Zhiyuan Wu, Sheng Sun, Yuwei Wang, Min Liu, Ke Xu, Wen Wang, Xuefeng Jiang, Bo Gao, and Jinda Lu. Fedcache: A knowledge cache-driven federated learning architecture for personalized edge intelligence. *IEEE Transactions on Mobile Computing*, 2024c.

Yunjia Xi, Weiwen Liu, Jianghao Lin, Jieming Zhu, Bo Chen, Ruiming Tang, Weinan Zhang, Rui Zhang, and Yong Yu. Towards open-world recommendation with knowledge augmentation from large language models. *arXiv preprint arXiv:2306.10933*, 2023.

Chunqiu Steven Xia, Yinlin Deng, Soren Dunn, and Lingming Zhang. Agentless: Demystifying llm-based software engineering agents. *arXiv preprint arXiv:2407.01489*, 2024.

Benfeng Xu, An Yang, Junyang Lin, Quan Wang, Chang Zhou, Yongdong Zhang, and Zhendong Mao. Expert-prompting: Instructing large language models to be distinguished experts. *arXiv preprint arXiv:2305.14688*, 2023.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing, 2024a.

Zihao Xu, Yi Liu, Gelei Deng, Yuekang Li, and Stjepan Picek. Llm jailbreak attack versus defense techniques–a comprehensive study. *arXiv preprint arXiv:2402.13457*, 2024b.

Biwei Yan, Kun Li, Minghui Xu, Yueyan Dong, Yue Zhang, Zhaochun Ren, and Xiuzheng Cheng. On protecting the data privacy of large language models (llms): A survey. *arXiv preprint arXiv:2403.05156*, 2024a.

Lixiang Yan, Lele Sha, Linxuan Zhao, Yuheng Li, Roberto Martinez-Maldonado, Guanliang Chen, Xinyu Li, Yueqiao Jin, and Dragan Gašević. Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, 55(1):90–112, 2024b.

Diyi Yang. *Computational Social Roles*. PhD thesis, Carnegie Mellon University Pittsburgh, PA, USA, 2019.

Diyi Yang and Lucie Flek. Towards user-centric text-to-text generation: A survey. In *Text, Speech, and Dialogue: 24th International Conference, TSD 2021, Olomouc, Czech Republic, September 6–9, 2021, Proceedings*, pp. 3–22, Berlin, Heidelberg, 2021. Springer-Verlag. ISBN 978-3-030-83526-2. doi: 10.1007/978-3-030-83527-9_1. URL https://doi.org/10.1007/978-3-030-83527-9_1.

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

Rui Yang, Xiaoman Pan, Feng Luo, Shuang Qiu, Han Zhong, Dong Yu, and Jianshu Chen. Rewards-in-context: Multi-objective alignment of foundation models with dynamic preference adjustment. *arXiv preprint arXiv:2402.10207*, 2024a.

Songhua Yang, Hanjie Zhao, Senbin Zhu, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and Hongying Zan. Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19368–19376, 2024b.

Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, pp. 100211, 2024.

Yelp. Yelp dataset challenge, 2014. URL https://www.yelp.com/dataset/challenge.

Bin Yin, Junjie Xie, Yu Qin, Zixiang Ding, Zhichao Feng, Xiang Li, and Wei Lin. Heterogeneous knowledge fusion: A novel approach for personalized recommendation via llm. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pp. 599–601, 2023.

Fangyi Yu, Lee Quartey, and Frank Schilder. Legal prompting: Teaching a language model to think like a lawyer. *arXiv preprint arXiv:2212.01326*, 2022.

Xiaohan Yu, Li Zhang, Xin Zhao, Yue Wang, and Zhongrui Ma. Ra-rec: An efficient id representation alignment framework for llm-based recommendation. *arXiv preprint arXiv:2402.04527*, 2024.

Jiayi Yuan, Ruixiang Tang, Xiaoqian Jiang, and Xia Hu. Large language models for healthcare data augmentation: An example on patient-trial matching. *AMIA Annu. Symp. Proc.*, 2023:1324–1333, 2023.

Xinfeng Yuan, Siyu Yuan, Yuhan Cui, Tianhe Lin, Xintao Wang, Rui Xu, Jiangjie Chen, and Deqing Yang. Evaluating character understanding of large language models via character profiling from fictional works. *arXiv preprint arXiv:2404.12726*, 2024.

Shengbin Yue, Wei Chen, Siyuan Wang, Bingxuan Li, Chenchen Shen, Shujun Liu, Yuxuan Zhou, Yao Xiao, Song Yun, Wei Lin, et al. Disc-lawllm: Fine-tuning large language models for intelligent legal services. *arXiv preprint arXiv:2309.11325*, 2023.

Hansi Zeng, Surya Kallumadi, Zaid Alibadi, Rodrigo Nogueira, and Hamed Zamani. A personalized dense retrieval framework for unified information access. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 121–130, 2023.

Qiusi Zhan, Zhixiang Liang, Zifan Ying, and Daniel Kang. InjecAgent: Benchmarking indirect prompt injections in tool-integrated large language model agents. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics ACL 2024*, pp. 10471–10506, Bangkok, Thailand and virtual meeting, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-acl.624.

Junjie Zhang, Ruobing Xie, Yupeng Hou, Wayne Xin Zhao, Leyu Lin, and Ji-Rong Wen. Recommendation as instruction following: A large language model empowered recommendation approach. *arXiv preprint arXiv:2305.07001*, 2023a.

Kai Zhang, Fubang Zhao, Yangyang Kang, and Xiaozhong Liu. Memory-augmented llm personalization with short- and long-term memory coordination, 2023b.

Kai Zhang, Yangyang Kang, Fubang Zhao, and Xiaozhong Liu. LLM-based medical assistant personalization with short- and long-term memory coordination. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2386–2398, Mexico City, Mexico, June 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.132. URL https://aclanthology.org/2024.naacl-long.132.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*, 2023c.

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

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

Yuntong Zhang, Haifeng Ruan, Zhiyu Fan, and Abhik Roychoudhury. Autocoderover: Autonomous program improvement. *arXiv preprint arXiv:2404.05427*, 2024b.

Zhehao Zhang, Jiaao Chen, and Diyi Yang. Mitigating biases in hate speech detection from a causal perspective. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 6610–6625, Singapore, December 2023d. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.440. URL https://aclanthology.org/2023.findings-emnlp.440.

Zhehao Zhang, Jiaao Chen, and Diyi Yang. Darg: Dynamic evaluation of large language models via adaptive reasoning graph. *arXiv preprint arXiv:2406.17271*, 2024c.

Jujia Zhao, Wenjie Wang, Chen Xu, Zhaochun Ren, See-Kiong Ng, and Tat-Seng Chua. Llm-based federated recommendation. *arXiv preprint arXiv:2402.09959*, 2024a.

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

Zihuai Zhao, Wenqi Fan, Jiatong Li, Yunqing Liu, Xiaowei Mei, Yiqi Wang, Zhen Wen, Fei Wang, Xiangyu Zhao, Jiliang Tang, et al. Recommender systems in the era of large language models (llms). *IEEE Transactions on Knowledge and Data Engineering*, 2024b.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023a.

Xiaosen Zheng, Tianyu Pang, Chao Du, Qian Liu, Jing Jiang, and Min Lin. Cheating automatic llm benchmarks: Null models achieve high win rates. *arXiv preprint arXiv:2410.07137*, 2024.

Zhi Zheng, Zhaopeng Qiu, Xiao Hu, Likang Wu, Hengshu Zhu, and Hui Xiong. Generative job recommendations with large language model. *arXiv preprint arXiv:2307.02157*, 2023b.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36, 2024a.

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Libiao Peng, Jiaming Yang, Xiyao Xiao, et al. Characterglm: Customizing chinese conversational ai characters with large language models. *arXiv preprint arXiv:2311.16832*, 2023.

Yujia Zhou, Zhicheng Dou, Bingzheng Wei, Ruobing Xie, and Ji-Rong Wen. Group based personalized search by integrating search behaviour and friend network. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 92–101, 2021.

Yujia Zhou, Qiannan Zhu, Jiajie Jin, and Zhicheng Dou. Cognitive personalized search integrating large language models with an efficient memory mechanism. In *Proceedings of the ACM on Web Conference 2024*, WWW '24, pp. 1464–1473, New York, NY, USA, 2024b. Association for Computing Machinery. ISBN 9798400701719. doi: 10.1145/3589334.3645482. URL https://doi.org/10.1145/3589334.3645482.

Banghua Zhu, Jiantao Jiao, and Michael Jordan. Principled reinforcement learning with human feedback from pairwise or K-wise comparisons. *CoRR*, abs/2301.11270, 2023a. URL https://arxiv.org/abs/2301.11270.

Kaijie Zhu, Jiaao Chen, Jindong Wang, Neil Zhenqiang Gong, Diyi Yang, and Xing Xie. Dyval: Graph-informed dynamic evaluation of large language models. *arXiv preprint arXiv:2309.17167*, 2023b.

Yuchen Zhuang, Haotian Sun, Yue Yu, Qifan Wang, Chao Zhang, and Bo Dai. Hydra: Model factorization framework for black-box llm personalization. *arXiv preprint arXiv:2406.02888*, 2024.

Mingchen Zhuge, Changsheng Zhao, Dylan Ashley, Wenyi Wang, Dmitrii Khizbullin, Yunyang Xiong, Zechun Liu, Ernie Chang, Raghuraman Krishnamoorthi, Yuandong Tian, et al. Agent-as-a-judge: Evaluate agents with agents. *arXiv preprint arXiv:2410.10934*, 2024.

Jingming Zhuo, Songyang Zhang, Xinyu Fang, Haodong Duan, Dahua Lin, and Kai Chen. ProSA: Assessing and understanding the prompt sensitivity of LLMs. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 1950–1976, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.108. URL https://aclanthology.org/2024.findings-emnlp.108/.

Cai-Nicolas Ziegler, Sean M McNee, Joseph A Konstan, and Georg Lausen. Improving recommendation lists through topic diversification. In *Proceedings of the 14th international conference on World Wide Web*, pp. 22–32, 2005.

Caleb Ziems, William Held, Jingfeng Yang, Jwala Dhamala, Rahul Gupta, and Diyi Yang. Multi-VALUE: A framework for cross-dialectal English NLP. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 744–768, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.44. URL https://aclanthology.org/2023.acl-long.44/.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. Can Large Language Models Transform Computational Social Science? *Computational Linguistics*, 50(1):237–291, 03 2024. ISSN 0891-2017. doi: 10.1162/coli_a_00502. URL https://doi.org/10.1162/coli_a_00502.

Thomas P. Zollo, Andrew Wei Tung Siah, Naimeng Ye, Ang Li, and Hongseok Namkoong. Personalllm: Tailoring llms to individual preferences. *arXiv preprint arXiv:2409.20296*, 2024.