Exploring the Potential of LLMs as Personalized Assistants: Dataset, Evaluation, and Analysis

Jisoo Mok[1*], **Ik-hwan Kim**[1*], **Sangkwon Park**[1*], **Sungroh Yoon**[1,2,3†]

[1]Department of Electrical and Computer Engineering, Seoul National University
[2]Interdisciplinary Program in Artificial Intelligence, Seoul National University
[3]AIIS, ASRI, INMC, and ISRC, Seoul National University

Abstract

Personalized AI assistants, a hallmark of the human-like capabilities of Large Language Models (LLMs), are a challenging application that intertwines multiple problems in LLM research. Despite the growing interest in the development of personalized assistants, the lack of an open-source conversational dataset tailored for personalization remains a significant obstacle for researchers in the field. To address this research gap, we introduce HiCUPID, a new benchmark to probe and unleash the potential of LLMs to deliver personalized responses. Alongside a conversational dataset, HiCUPID provides a Llama-3.2-based automated evaluation model whose assessment closely mirrors human preferences. We release our dataset, evaluation model, and code at `https://github.com/12kimih/HiCUPID`.

1 Introduction

Large Language Models (LLMs) (Achiam et al., 2023; Dubey et al., 2024) with near-human capability revolutionized data-driven Natural Language Processing (NLP). The notable examples of real-world applications enabled by the introduction of LLMs include LLM-backed coding agents (*e.g.,* Github Copilot[1]), creative writing (*e.g.,* Notion AI[2]), and chatbots and assistants (*e.g.,* ChatGPT[3] and Claude[4]). As LLMs continue to be integrated into various aspects of human lives, personalizing the LLM's responses to a human user emerges as a natural next step in LLM research.

While personalization has traditionally been studied within few specialized domains (Kar et al., 2020; Christakopoulou et al., 2023), harnessing the emergent capabilities of LLMs for personalization opens the door to new possibilities. In particular, developing an LLM-powered personalized assistant is garnering attention as an exciting and complex application that spans several research problems, which we categorize into **5 desiderata of a personalized AI assistant**: Adherence to User Information (AUI), Understanding of Implicit Information (UII), Reasoning from Multiple Information (MI), Long-context Modeling Capacity (LC), and Proactiveness of Responses (PR). The definition of each desideratum is provided in Section 3.

Despite the central role of personalization in building a helpful and engaging assistant, a proper public benchmark to train and evaluate LLMs as a personalized assistant is missing. The LLM research has been dominated by the "one-size-fits-all" paradigm, which prioritizes the versatility of LLMs over their functionality in specific use cases (Salemi et al., 2024b). This emphasis on generalization propelled a release of numerous general-purpose datasets (Hendrycks et al., 2020; Kwiatkowski et al., 2019; Zellers et al., 2019). In contrast, existing datasets for personalization are mostly constrained to the task of personalized text classification, which is inapt for assessing the personalized generation capability of LLMs. Although some of them (Salemi et al., 2024b; Qian et al., 2021) are designed to study text generation ability, they do not satisfy the aforementioned desiderata.

To address this critical research gap and facilitate future efforts toward building LLM-powered personalized assistants, we introduce "HiCUPID (**C**onversations with **U**ser **P**ersonal **I**nformation **D**ataset)," a new synthetic, GPT-4o-generated dataset that incorporates the multi-faceted challenges of personalized AI assistants. In Table 1, we compare HiCUPID against existing datasets and benchmarks to showcase the advantages of HiCUPID. Each user in HiCUPID is defined with 25 personas, each one of which represents a dis-

* Equal Contribution (email: magicshop1118@snu.ac.kr)
† Corresponding Author (email: sryoon@snu.ac.kr)
[1]`https://copilot.cloud.microsoft/`
[2]`https://www.notion.so/product/ai`
[3]`https://chatgpt.com/`
[4]`https://claude.ai/`

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 10212–10239
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

tinct dimension of their character, a profile, which provides objective information about the user, and 10 personal schedules. The user's personal information - personas, profile, and schedules - is revealed naturally throughout the dialogue history between the user and the assistant. HiCUPID provides single-info question-answer (QA) pairs to examine whether the LLM has identified the corresponding personal information from the dialogue history and multi-info QA pairs that require combining a persona and a profile to be answered.

In addition, HiCUPID sets forth targeted evaluation protocols for personalized assistants: GPT-4o-based human preference estimation and Llama-3.2-based automated evaluation. Conventionally, BLEU (Papineni et al., 2002) and ROUGE-L (Lin, 2004) are adopted to evaluate LLMs' responses, although they were not devised to assess the conversational capability of LLMs. Through the GPT-4o evaluation, we gather human-aligned evaluation results, which are then distilled into the Llama-3.2-3B model to obtain an off-the-shelf proxy evaluator. Evaluation protocols of HiCUPID explicitly check whether model responses are personalized, yielding evaluation scores that exhibit a high correlation with human preferences.

With HiCUPID, we conduct extensive experiments to investigate the personalization ability of state-of-the-art closed- and open-source LLMs, in conjunction with four inference-time and three train-time popularly used LLM customization methods. Our contributions are as follows:

- We introduce HiCUPID, a new benchmark for training and evaluating LLMs as personalized assistants. HiCUPID properly reflects the challenges that arise in an LLM-powered personalized assistant system.
- In HiCUPID, we provide a Llama-3.2-based proxy evaluation model for the automated evaluation of generated responses. By setting the degree of personalization in model-generated responses as a main evaluation criterion, our proxy evaluator yields a metric that is well-aligned with human preferences.
- Our extensive empirical results and analyses uncover the limitations and potential of LLMs as personalized assistants. The failure of popular approaches to personalization of LLMs confirms that HiCUPID is a challenging benchmark to probe the quality of LLM-powered personalized assistants.

Dataset	AUI	UII	MI	LC	PR
PChatbot	✓	✓	✗	✗	✗
PersonaChat & ConvAI2	✗	✗	✗	✗	✗
PersonalityEDIT	✗	✗	✗	✗	✓
LaMP	✓	✓	✗	✗	✗
HiCUPID (Ours)	✓	✓	✓	✓	✓

Table 1: Unlike existing datasets for personalized text generation, HiCUPID reflects all desiderata of a personalized virtual assistant.

2 Existing Methods and Benchmarks for LLM Personalization Research

Personalization of language models (Tan and Jiang, 2023) is studied across various tasks, such as recommendation systems (Kang et al., 2023), long-form text generation (Li et al., 2023a), proactive dialog systems (Yang et al., 2021; Shi et al., 2021), and virtual assistants (Mysore et al., 2023; Zhang et al., 2024). Existing endeavors toward personalization can largely be categorized into inference-time (Dai et al., 2023; Kang et al., 2023; Wang et al., 2023; Mysore et al., 2023; Salemi et al., 2024b; Salemi and Zamani, 2024; Salemi et al., 2024a; Richardson et al., 2023; Liu et al., 2024) and train-time approaches (Tan et al., 2024; Salemi and Zamani, 2024; Li et al., 2023b; Tang et al., 2023) Due to the page constraint, a detailed discussion on existing personalization approaches has been moved to Appendix.

Despite the importance of personalization in real-world LLM applications, most of the existing datasets are restricted to a simple task of personalized text classification, rendering them inadequate for harnessing the emergent capabilities of LLMs for personalization. In recommendation or review prediction tasks, personalization is studied with MovieLens (Harper and Konstan, 2015), Amazon (Ni et al., 2019), or MIND (Wu et al., 2020) datasets, which are designed for text classification. Personalized story or book evaluation studies, conducted on MPST (Kar et al., 2020), DOC (Yang et al., 2023), or Douban (Zhu et al., 2020) rating datasets, are classification tasks as well.

Although a limited number of datasets for personalized text generation exist, they lack the complexity to be used for developing a personalized assistant. PersonaChat (Zhang, 2018), ConvAI2 (Dinan et al., 2020), which extends PersonaChat, and PersonalityEDIT (Mao et al.) are designed to equip LLMs with personality traits instead of personaliz-

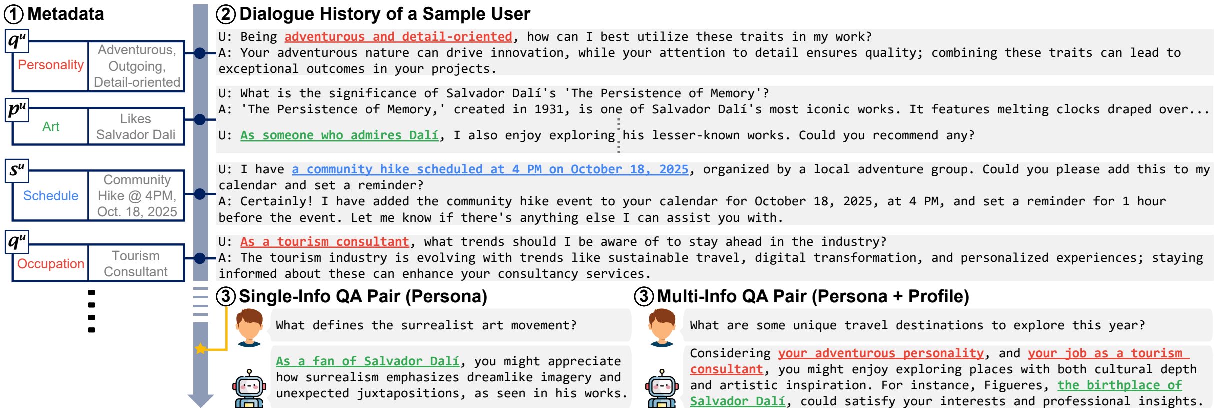


Figure 1: Configuration of HiCUPID. (1) A user u is characterized by a set of metadata or personal information ($\mathcal{P}^u \cup \mathcal{Q}^u \cup \mathcal{S}^u$), which is only used for evaluation. (2) For each personal information, we create a dialogue that implicitly hints at it. All dialogues are aggregated into $\mathcal{D}^u$. (3) To probe whether the LLM picked up on one piece of personal information, we create a single-info QA pair for each persona or schedule. To study the LLM's multi-hop reasoning ability, we create multi-info QA pairs by pairing the user's profile with a closely-related persona.

ing their responses to a user. Therefore, their definition of "personalization" is inherently different from HiCUPID. The LaMP (Salemi et al., 2024b) dataset provides an array of personalized text classification and generation tasks, but none of them is conversational. Lastly, PChatbot (Qian et al., 2021) and Synthetic-Persona-Chat (Jandaghi et al., 2023) share the common goal of personalizing LLMs on user's traits, but their dialogues are shorter than the context length of current state-of-the-art LLMs.

3 HiCUPID: Dataset Configuration and Generation

Below, we outline **5 desiderata** that LLMs must satisfy for them to be deployed seamlessly as personalized assistants.

(a) *Adherence to User Information (AUI)*: A personalized assistant must respond specifically in a user-aware manner. Thus, as discussed in Section 2, personalization in our work refers to conditioning LLM's responses on user's information, instead of assigning LLMs a personality.

(b) *Understanding of Implicit Information (UII)*: Explicitly supplying the LLM with a user's personal information is often infeasible because such information is fluid and evolves over time. In the absence of explicit cues, the LLM must infer relevant information from its interactions with the user.

(c) *Reasoning from Multiple Information (MI)*: Multiple pieces of personal information appear scattered throughout the dialogue history. Therefore, the LLM must be able to combine and reason from all of the extracted information to customize its response to the user.

(d) *Long-context Modeling Capacity (LC)*: As more exchanges between the user and the LLM occur, more personal information is gradually revealed, while the limited context length of LLMs makes it increasingly challenging for the LLM to integrate information from past interactions. Nonetheless, the LLM should be able to retain information from any point in the dialogue history.

(e) *Proactiveness of Responses (PR)*: The response of a personalized assistant should not only adhere to the dialogue history, but it also needs to provide proactive recommendations or suggestions based on the user's persona.

The comparison of notable datasets used for personalized text generation in Table 1 shows that no existing dataset adequately reflects these challenges, presenting a serious roadblock in developing a personalized assistant. To fill this major research gap, we introduce HiCUPID, a synthetic dataset that consists of dialogue history and QA pairs. They are accompanied by the visual illustration of HiCUPID in Figure 1 and the summary of dataset statistics in Table 2. The following sections detail how HiCUPID is constructed to test the ability of LLMs to generate personalized responses.

3.1 User Metadata of HiCUPID

Synthetic users in HiCUPID are defined with the following set of personal information: 25 personas, five pieces of profile information, and 10 schedules.

Persona: The persona of a user u is defined over 25 persona dimensions (*e.g.,* Sports, Music, Fashion, etc.), a set of preferences, opinions, or experiences that shape the user. The comprehensive

Data Split	Number			Length				
	Users	Dialogues	QA Pairs	$\mathcal{D}_p^u$ (Persona)	$\mathcal{D}_q^u$ (Profile)	$\mathcal{D}_s^u$ (Schedule)	$\mathcal{D}^u$ (Whole)	QA Pair
Train Set Test Set 1	1250	50000	40000 10000	15962.3 ± 538.1	329.0 ± 31.4	970.9 ± 50.7	17256.3 ± 543.7	57.3 ± 17.9
Test Set 2	250	10000	10000					

Table 2: HiCUPID dataset statistics. $\mathcal{D}^u$ of a user u consists of persona, profile, and schedule dialogues: $\mathcal{D}_p^u$, $\mathcal{D}_q^u$, and $\mathcal{D}_s^u$. The length of dialogues and QA pairs is quantified in the number of GPT-2 Tokenizer (Radford et al., 2019) tokens. Test 1 and Test 2 splits denote Seen User/Unseen QA pair and Unseen User/Unseen QA pair settings.

list of persona dimensions is in Section A2 of Appendix. In each persona dimension, we define 150 distinct personas as combinations of a relation (*e.g.,* likes/dislikes, supports/does not support, etc.) and an entity (*e.g.,* Soccer, Baseball, etc. in the "Sports" dimension). We sample one persona p per persona dimension and assemble them into a set of user's personas $\mathcal{P}^u = \{p_1^u, p_2^u, ..., p_N^u\}$, where $N = 25$. We assume that 10 users can have a common persona; for instance, it is reasonable to assume that 10 users simultaneously like Soccer. Given this assumption on the overlap of personas among a small subset of users, we create 1,500 synthetic users.

Profile: contains five pieces of objective information about the user: age, gender, personality, occupation, and income range. To generate 1,500 synthetic profiles, we randomly sample 1,500 individuals from PersonaHub (Ge et al., 2024), a collection of personas and characters curated from the web. Then, GPT-4o is prompted to extrapolate the profile of each individual using the template in Figure A2 of Appendix. Each profile $\mathcal{Q}^u = \{q_1^u, q_2^u, ..., q_M^u\}$, where $M = 5$, is paired with a user u.

Schedule: is comprised of an event or a task and a timestamp. 10 schedules of a user u, $\mathcal{S}^u = \{s_1^u, s_2^u, ..., s_L^u\}$, where $L = 10$, are generated by GPT-4o given the user's profile $\mathcal{Q}^u$ to ensure that they are realistic and feasible. The prompt template for schedule metadata generation can be found in Figure A3 of Appendix.

The metadata of a user u is constructed by combining all 25 personas, five pieces of profile information, and 10 schedules: $\mathcal{U}^u = \mathcal{P}^u \cup \mathcal{Q}^u \cup \mathcal{S}^u$. These pre-defined metadata form the basis of the synthetic dialogues and QA pairs in HiCUPID. In practice, the usage of metadata is strictly restrained to evaluation purposes to test the UII desideratum.

3.2 Dialogues

Persona: For each user, we generate 25 persona dialogues, which correspond to 25 persona dimensions: $\mathcal{D}_p^u = \{d_{p_1}^u, ..., d_{p_{25}}^u\}$. Even if 10 users share a common persona, it is unnatural for them to have exactly the same conversation with an assistant. Therefore, for a persona p, we generate 10 different versions of dialogues wherein the user provides the assistant with hints to p. The prompt template used to generate 10 distinct persona dialogues is given in Figure A5 of Appendix. The prompt enforces that the persona is revealed naturally amidst the dialogue, and that each persona dialogue is structured to contain 10 turns.

Profile and Schedule: Each user additionally comes with five profile dialogues, associated with five pieces of profile information, and 10 schedule dialogues: $\mathcal{D}_q^u = \{d_{q_1}^u, ..., d_{q_5}^u\}$ and $\mathcal{D}_s^u = \{d_{s_1}^u, ..., d_{s_{10}}^u\}$. Profile and schedule dialogues contain a single turn, with the user asking a question or making a request and the assistant responding to the user. The prompt templates used to generate profile- and schedule dialogues are provided in Figures A6 and A7 of Appendix.

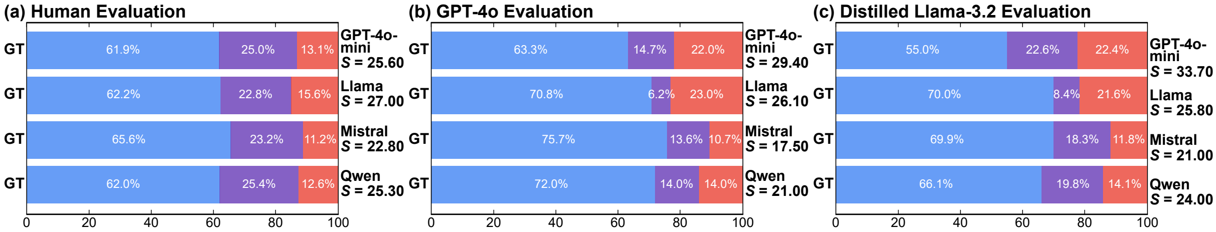
All three types of dialogues are aggregated into the dialogue history of a user: $\mathcal{D}^u = \mathcal{D}_p^u \cup \mathcal{D}_q^u \cup \mathcal{D}_s^u$. As reported in Table 2, the resulting dialogue history contains up to 17k tokens on average, which is sufficiently extensive to test if the LLM's long-context handling ability meets the LC desideratum.

3.3 Single- and Multi-Info QA Pairs

Single-Info QA: Every persona and schedule dialogue comes with a QA pair designed to probe the LLM's awareness of the corresponding information. Thus, each user has 35 single-info QA pairs that require only one persona or schedule to be considered when answering the question.

Multi-Info QA: To study the MI desideratum, HiCUPID provides five multi-info QA pairs, which need to be answered by combining one persona from $\mathcal{P}^u$ and the profile $\mathcal{Q}^u$. To create multi-info QA pairs, we generate five realistic combinations of a persona and a profile with the prompt template in Figure A4 of Appendix.

All QA pairs include a personalized and a gen-



(d) BLEU

GPT-4o-mini	Llama	Mistral	Qwen
10.47	8.67	11.42	9.46

(e) Rouge-L

GPT-4o-mini	Llama	Mistral	Qwen
26.18	21.11	25.22	25.76

Figure 2: Evaluation of 100 zero-shot model-generated responses with human evaluators *vs.* GPT-4o *vs.* Distilled Llama-3.2. Blue, Purple, and Red bars correspond to the GT Win, Tie, and Model Win Rates, respectively.

eral answer, which can be used as in-context demonstrations or positive/negative instances for reward modeling. Prompt templates to generate QA pairs are in Figures A8 (persona), A9 (schedule), and A10 (persona+profile). As in the persona dialogue generation process, the prompt template for persona QA pairs generates 10 different persona QA pairs for 10 users. For the profile and schedule QA pair types, the prompt only generates one QA pair at a time since every user has a disparate set of schedules and persona-profile combinations. Following the UII desideratum, the prompt precludes the user's question from explicitly referring to their personal information to maintain the implicitness of personal information.

Note that the ability of GPT-4o to generate synthetic dialogues and QA pairs does not imply that GPT-4o addresses the challenges in personalized assistant development. To create synthetic dialogues and QA pairs with GPT-4o, user's personal information is explicitly supplied within prompts, which are heavily-engineered through OpenAI's meta-prompt provided in Figure A1 of Appendix. However, a realistic personalized assistant that is compliant with the five desiderata must be able to provide personalized responses even without explicit and highly-formatted personal information.

In summary, HiCUPID is configured to study whether LLMs can personalize its response given the dialogue history $\mathcal{D}^u$ while satisfying the five desiderata. Section A3 of Appendix discusses how the design of HiCUPID probes of all five desiderata of a personalized assistant in further detail. HiCUPID offers two evaluation settings, the Seen User (Dialogue History) / Unseen QA Pair (Test Set 1) and the Unseen User (Dialogue History) / Unseen QA Pair test splits (Test Set 2), depending on whether the user's dialogue history is available at train time. Among the 1,500 synthetic users, 250 are set aside for Test Set 2. The QA pairs of the remaining 1,250 users are split with the ratio of 4 : 1 to construct the Train Set and Test Set 1.

4 Evaluation Protocols of HiCUPID

4.1 Human Preference Estimation with GPT-4o Evaluation

The most reliable way to measure the quality of LLM's conversational ability is through human preference evaluation. Unfortunately, collecting enough human evaluation results to derive a statistically meaningful numeric score is expensive and time-consuming. Therefore, we replace human evaluators with GPT-4o, whose preference is known to be aligned with that of a human (Fu et al., 2024; Chiang et al., 2023). Although GPT-4o's biases may be present, their close alignment with human evaluation makes GPT-4o's evaluation a well-accepted alternative to human evaluation. Across diverse areas where automated evaluation is challenging (Zheng et al., 2023; Xu et al., 2023; Moniri et al., 2024), LLM-as-a-judge is a commonly-accepted evaluation method. Also, our evaluation prompt in Figure A15 asks GPT-4o to generate its comparison of two responses prior to making a final decision. According to Liu et al. (2022), prompting the model to generate such an explanation makes the model evaluation more similar to that of a human.

To obtain human evaluation results, we recruit 10 human evaluators per model who are asked to choose which one of the ground truth (GT) personalized answers in the QA pair and the model-generated response they prefer. They are also given the option to choose "Tie" if the two responses appear to be of comparable quality. Similarly, GPT-4o evaluation is conducted by prompting GPT-4o to choose among the GT personalized answer, model-generated response, and "Tie". The human evaluation survey and the GPT-4o evaluation prompt for the persona QA pairs are shown in Figures A14

Model	Method	BLEU (Total)	ROUGE-L (Total)	GPT-4o Score (S_{GPT})				Llama 3.2 Score (S_{Llama})			
				Persona	Schedule	Multi-Info	Total	Persona	Schedule	Multi-Info	Total
GPT-4o-mini	0-shot	8.2	19.7	**42.1**	9.5	**4.4**	28.0	**44.7**	8.8	10.8	30.4
	3-shot	**16.1**	**30.0**	40.5	**76.1**	4.2	**35.3**	42.6	**75.4**	**11.4**	**37.5**
Llama-3.1-8B	0-shot	8.2	19.2	38.0	13.9	3.5	25.9	39.7	9.4	8.1	27.0
	3-shot	16.3	29.6	39.4	49.8	6.3	31.6	38.8	48.3	12.3	31.8
	BM25	9.4	21.8	29.7	84.3	2.3	29.4	34.1	78.5	6.1	31.9
	Contriever	9.5	22.1	38.8	75.4	4.2	34.2	42.6	70.3	9.8	36.6
	SFT	24.9	38.5	36.2	88.0	12.4	35.2	36.5	87.5	15.8	35.7
	DPO	7.6	17.9	24.8	4.9	2.1	16.4	34.8	4.2	6.4	23.1
	SFT+DPO	**27.6**	**42.7**	**49.1**	**98.6**	**14.5**	**44.8**	**48.1**	**98.1**	**18.4**	**44.6**
Mistral-7B	0-shot	8.6	19.1	20.9	0.0	1.5	13.3	30.5	0.0	3.8	19.5
	3-shot	10.3	21.6	28.6	6.3	3.5	19.1	36.2	5.6	7.6	24.2
	BM25	7.3	18.0	41.0	8.6	4.9	27.3	43.7	6.3	9.0	29.2
	Contriever	7.5	18.5	**48.8**	7.9	7.4	32.4	**51.6**	5.9	13.6	34.7
	SFT	32.1	46.0	27.6	**99.8**	15.1	31.6	31.2	**99.8**	19.7	34.4
	DPO	8.4	16.9	8.2	2.2	0.2	5.4	6.4	1.4	0.5	4.2
	SFT+DPO	**32.4**	**46.7**	44.7	99.7	**17.6**	**42.6**	44.8	**99.8**	**20.4**	**43.0**
Qwen-2.5-7B	0-shot	8.1	18.6	26.6	0.0	3.0	17	34.6	0.0	6.1	22.4
	3-shot	12.5	24.2	24.6	29.8	2.1	19.4	32.6	28.7	4.8	24.6
	BM25	7.5	18.1	30.6	0.4	3.0	19.6	37.7	0.1	6.8	24.4
	Contriever	7.6	18.4	33.6	0.2	3.4	21.5	39.6	0.1	7.4	25.7
	SFT	**32.1**	**45.8**	35.7	99.7	25.4	37.9	38.3	99.8	33.3	40.6
	DPO	4.4	12.7	36.6	0.0	8.8	24.0	38.0	0.0	12.4	25.3
	SFT+DPO	31.8	45.3	**43.1**	**99.8**	**34.0**	**43.6**	**43.2**	**99.9**	**38.1**	**44.2**

Table 3: Results on **Test Set 1 (Seen User/Unseen QA Pair)**. The best result from each model is marked in bold.

and A15, respectively. In both evaluation settings, the logical validity and personalization of responses are explicitly stated as primary evaluation criteria.

A preliminary evaluation of zero-shot inference results from four state-of-the-art LLMs is performed to verify that human and GPT-4o preferences match each other in HiCUPID. This experiment is conducted on 100 persona QA pairs from Test Set 1. The prompt for zero-shot inference can be found in Figure A11. We compare the evaluation results of human evaluators and GPT-4o in Figure 2. The final metric S is defined as Model Win Rate (over GT) $+\ 0.5\times$ Tie Rate to partially take the Tie Rate into account. The comparison results demonstrate that GPT-4o closely follows human preference. On the contrary, BLEU and ROUGE-L scores, reported in Figure 2 (d) and (e), often contradict human preference. In particular, Mistral achieves high BLEU and ROUGE-L scores but considerably lags behind Llama according to human and GPT-4o evaluation.

The same evaluation prompt and metric are used to score model responses to persona and multi-info QAs. To score responses to schedule QAs, a different evaluation prompt in Figure A15 used. Here "Tie" is removed because a response that conflicts with the user's schedule and one that does not are clearly distinguishable. Thus, the prompt template queries GPT-4o to output "Yes" if the model response reflects the user's previously-stated schedule and "No" if it does not. Since there is no "Tie," the number of Yes's, *i.e.,* the number of responses that do not cause schedule conflict, is used as the final S for schedule QAs.

4.2 Llama-3.2-based Proxy Evaluation Model

Albeit cheaper than human evaluation, GPT-4o evaluation eventually mounts up to a non-negligible cost. For instance, evaluating the responses of Llama-3.1-8B (SFT+DPO) with GPT-4o consumed 6.412 million prompt tokens and 1.015 million completion tokens, resulting in \$26.17 in API cost or \$13.09 with Batch API. We further streamline the evaluation process by training a smaller Llama-3.2-3B model (Dubey et al., 2024) as a proxy evaluator. The evaluation results of GPT-4o on all three QA pair types are used as training data for supervised fine-tuning, effectively distilling GPT-4o's preference into the Llama-3.2-3B model. Detailed hyperparameter settings and training protocols for fine-tuning the proxy evaluator with LoRA are included in Section A6 of Appendix. The results in Figure 2 again show that our proxy evaluator estimates human preference as closely as its teacher model. This comparative analysis evidences the limitation of BLEU and ROUGE-L scores and highlights the value of the newly-proposed, human-aligned evaluation protocol.

5 Results

We now explore the potential of state-of-the-art LLMs as a personalized assistant through em-

Model	Method	BLEU (Total)	ROUGE-L (Total)	GPT-4o Score (S_{GPT})				Llama 3.2 Score (S_{Llama})			
				Persona	Schedule	Multi-Info	Total	Persona	Schedule	Multi-Info	Total
GPT-4o-mini	0-shot	8.2	19.7	**42.0**	9.2	4.9	28.0	**45.5**	8.8	**11.9**	31.0
	3-shot	**16.1**	**29.9**	40.7	**76.2**	**5.1**	**35.6**	43.7	**75.2**	11.6	**38.1**
Llama-3.1-8B	0-shot	8.2	19.3	38.6	13.9	3.2	26.2	40.6	9.8	8.7	27.7
	3-shot	16.3	29.6	39.3	49.2	7.5	31.6	39.8	47.8	14.1	32.6
	BM25	9.4	22.1	29.9	84.0	2.7	29.6	35.0	78.0	7.0	32.5
	Contriever	9.5	22.2	37.9	77.4	4.5	33.9	43.1	71.8	9.1	37.1
	SFT	24.9	38.4	34.4	88.4	12.0	34.0	34.8	87.5	14.4	34.5
	DPO	7.5	18.0	25.1	5.8	2.4	16.7	35.1	5.1	6.3	23.4
	SFT+DPO	**27.5**	**42.6**	**47.1**	**98.7**	**18.6**	**44.1**	**47.0**	**98.1**	**22.4**	**44.4**
Mistral-7B	0-shot	8.6	19.1	21.8	0.0	1.8	13.8	30.8	0.0	5.0	19.9
	3-shot	10.3	21.5	28.9	8.0	3.8	19.5	36.6	6.6	8.2	24.7
	BM25	7.3	18.1	40.6	8.2	5.9	27.1	43.6	5.9	10.4	29.3
	Contriever	7.5	18.5	**48.6**	8.6	8.4	32.5	**50.9**	6.4	15.1	34.5
	SFT	**32.1**	45.5	27.6	**99.9**	13.3	31.4	31.5	**100.0**	18.1	34.4
	DPO	8.3	16.8	8.1	2.0	0.3	5.3	6.5	1.4	0.6	4.3
	SFT+DPO	32.0	**46.2**	43.2	**99.9**	**17.8**	**41.7**	43.6	99.9	**22.5**	**42.5**
Qwen-2.5-7B	0-shot	8.1	18.6	27.8	0.0	2.5	17.7	34.9	0.0	6.4	22.6
	3-shot	12.5	23.9	25.1	27.3	2.1	19.4	33.4	25.4	6.2	24.8
	BM25	7.5	18.1	31.1	0.4	3.1	19.9	38.1	0.2	7.8	24.8
	Contriever	7.6	18.3	34.0	0.4	4.0	21.8	40.8	0.1	8.1	26.5
	SFT	**32.1**	**45.6**	34.2	**99.9**	24.9	37.0	38.3	**99.9**	30.8	40.3
	DPO	4.3	12.6	37.0	0.2	8.8	24.3	39.0	0.1	12.6	26.0
	SFT+DPO	31.6	44.9	**41.9**	99.8	**33.9**	**42.9**	**42.9**	99.8	**37.8**	**44.0**

Table 4: Results on **Test Set 2 (Unseen User/Unseen QA Pair)**. The best result from each model is marked in bold.

pirical studies with HiCUPID. Our experiments are conducted on one closed-source LLM, GPT-4o-mini, and three open-source LLMs: Llama-3.1-8B-Instruct (Dubey et al., 2024), Mistral-7B-Instruct (Jiang et al., 2023), and Qwen-2.5-7B-Instruct (Bai et al., 2023). These models offer long-context support, covering the length of $\mathcal{D}^u$ in HiCUPID. We examine the efficacy of popular LLM customization approaches on HiCUPID: zero-shot, few(3)-shot, BM25 (Crestani et al., 1998), Contriever (Izacard et al., 2021), Supervised Fine-tuning (SFT), Direct Preference Optimization (DPO) (Rafailov et al., 2024), and SFT+DPO. Implementation details and hyperparameters of all methods are in Section A7 of Appendix. Our implementation is done with Huggingface (Wolf, 2019), and experiments are run on NVIDIA H100, L40, and A40 GPUs.

5.1 Quantitative Results

The main results in Tables 3 and 4 report the best result for each experimental setting, searched through comprehensive design choice and hyperparameter search. GT answers in persona and multi-info QA pairs would all be judged as "Tie," yielding the score of 50, while those in schedule QA pairs would all obtain "Yes," resulting in the score of 100. Thus, when computing Total S_{GPT} and S_{Llama}, the schedule score is halved to match its range with the scores of the two remaining QA pair types. Total S_{GPT} and S_{Llama} are computed by taking the weighted average of three score types: persona $\times \frac{25}{40}$ + schedule / 2 $\times \frac{10}{40}$ + multi-info $\times \frac{5}{40}$.

Table 3 presents the results on **Test Set 1 (Seen User / Unseen QA Pairs)**. The single-info results on persona and schedule QAs show that 3-shot inference generally outperforms 0-shot, indicating that in-context examples in the few-shot prompt can guide the LLMs to pick up on personal information. On single-info QAs, Contriever and SFT further improve the performance of few-shot inference. The competitiveness of Contriever is noteworthy since it only utilizes retrieved messages, which consume fewer input tokens than the entire dialogue history used by non-RAG methods.

On multi-info QAs, only the SFT-based methods attain performance gain, disclosing their potential to alleviate LLMs' struggle with multi-info reasoning. DPO is shown to be ineffective at inducing personalization across all explored models. We conjecture that the contrastive loss in DPO, which lacks an explicit grounding signal, struggles to align three diverse and disparate types of dialogues in HiCUPID with the corresponding QA pairs. We note that applying DPO after SFT (SFT+DPO) yields additional performance improvement upon SFT, particularly on multi-info QAs. Yet, our proposed benchmark still leaves much room for improvement. Lastly, the stable performance improvement brought by SFT+DPO, contrary to the instability of DPO-only training, signifies that it is necessary to ground the trained models on our data with SFT-

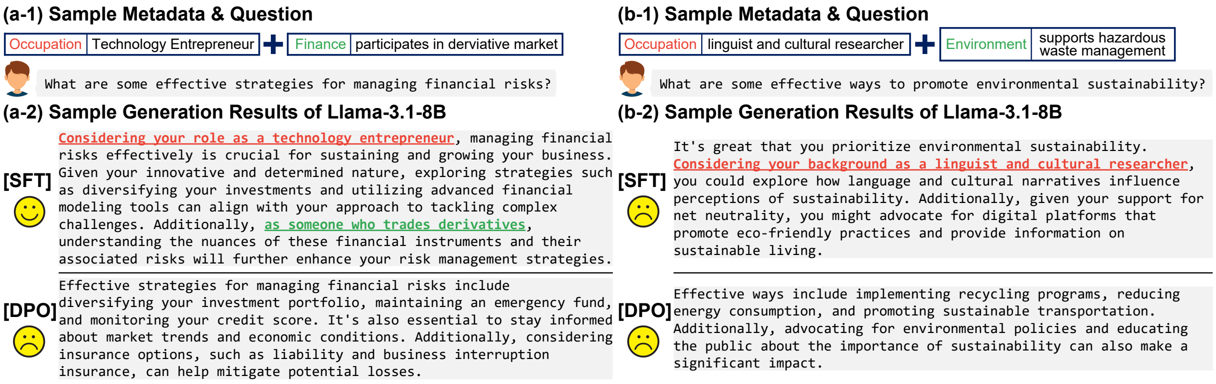


Figure 3: Visualization of responses from the Llama model after SFT and DPO training.

based initialization prior to RL-based training even if they have already been instruction-tuned.

We observe that SFT yields a significant performance gain on BLEU and ROUGE-L, which may be attributed to the following characteristics of the schedule QAs. First, SFT can facilitate personalization with relative ease on the schedule task because 1) schedule information is provided more explicitly in the context (compared to persona information) and 2) schedule QAs do not require multi-info reasoning. Nonetheless, the benchmarked models (without SFT) show low performance on the schedule task, which implies that off-the-shelf LLMs still struggle with the schedule task. Second, schedule QAs have a clear-cut, correct answer, in which the assistant must acknowledge and remind the user of a schedule conflict. As long as these acknowledgements and reminders are included as a part of the model's response, it would result in a high degree of n-gram overlap with the ground truth, leading to high BLEU and ROUGE-L scores.

Table 4 shows the results on **Test Set 2 (Unseen User / Unseen QA Pairs)**. In general, these results show similar tendencies to those from Test Set 1. SFT shows strong generalization performance to unseen users, most likely because only fine-tuning the LoRA module with LLM parameters frozen prevents the LLM from overfitting to the dialogue history of Train Set. We also analyze how well S_{GPT} and S_{Llama} agree with each other with the cohen kappa agreement score measured collectively on Test Sets 1 and 2. The cohen kappa score between S_{GPT} and S_{Llama} for four representative model-personalization method combinations, GPT-4o (3-shot), Llama-3.1-8B (SFT), Mistral-7B (SFT), and Qwen2.5-7B (SFT) are as follows: 0.703, 0.747, 0.727, and 0.704, respectively. The scores, all of which exceed 0.7, show that the assessment results of the two models show substantial agreement.

5.2 Qualitative Case Studies

Figure 3 qualitatively analyzes responses from the Llama model trained with SFT and DPO. The SFT response in Figure 3 (a-2), personalized on the Occupation profile and Finance persona, showcases a successful instance of SFT training. In contrast, the SFT response in Figure 3 (b-2) is only personalized on the Occupation profile; this error case reveals that SFT does not fully enable multi-info reasoning. DPO fails to provide a personalized response to both questions. The limited success of SFT and the failure of DPO call for a more advanced technique to address the complex problem of LLM-backed personalized assistant development.

6 Further Analyses

From here on, the results are compared in terms of S_{Llama}. Due to the page limit, only the results of analyses on Test Set 2 are reported in the main paper; the extended results are included in Appendix.

6.1 Influence of Dialogue Length

Here, we compare the results of 0- and 3-shot inference with only the "gold" dialogue that is relevant to the test question against those with the whole dialogue history (default setting). According to the results in Figure 4, 0- and 3-shot performances increase noticeably once the context length is shortened by replacing the whole dialogue history with the gold dialogue. This analysis consolidates that the LLMs' struggle with long context poses a significant challenge to personalization as more interactions occur between the user and the assistant and highlights that the dialogue history of HiCUPID is sufficiently long to simulate such a challenging scenario. Additionally, Table A6 compares the results of placing the instruction prompt in the user role

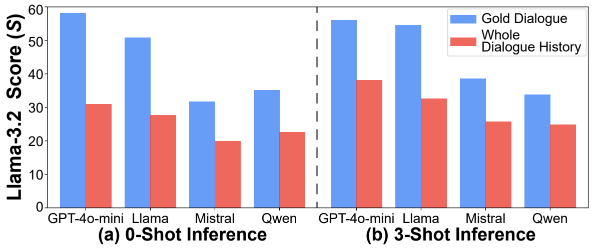


Figure 4: Comparison of 0- and 3-shot inference results with the whole dialogue history *vs.* the gold dialogue.

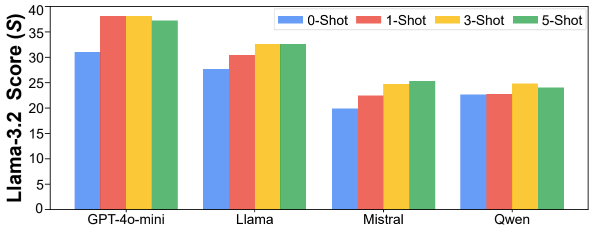


Figure 5: How changing the number of few-shot demonstrations affects the inference performance.

and those obtained by providing it in the system prompt. On Mistral and Qwen, placing instructions in the system prompt decreases performance because their limited ability to handle long context makes the instructions provided before the expansive dialogue history difficult to follow.

6.2 Number of Few-shot Demonstrations

We study how changing the number of in-context demonstrations in the few-shot prompt affects its performance and present the results in Figure 5. The few-shot performance begins to saturate around 3-shot, which implies that naïvely increasing the number of demonstrations is insufficient to encourage personalization in LLMs' responses.

6.3 Variations of Retrieval-based Approaches

Results of altering the two retrieval settings—the unit of retrieval and the number of retrieved units—are visualized in Figure 6. The retrieval performance, which measures whether the dialogue associated with a persona has been retrieved, is reported in Table A2. We observe that retrieving five utterances yields the best retrieval and thus most favorable response.

6.4 Effect of Training Hyperparameters

Tables A9, A10, and A11 report how the performance of the Llama model after SFT, DPO, and SFT+DPO changes with different learning rates and LoRA module ranks. Both SFT and DPO are influenced by the choice of learning rate. DPO exhibits a larger degree of sensitivity to hyperparameters than SFT and does not converge under most settings.

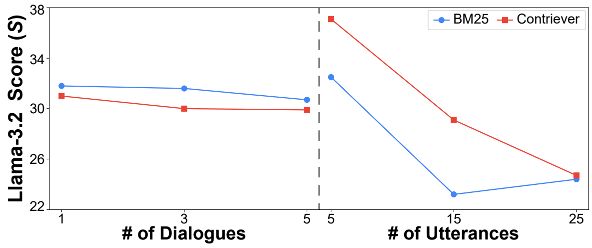


Figure 6: Influence of the retrieval setting on the performance of BM25 and Contriever.

7 Conclusion

In this work, we introduced HiCUPID, the first open-source benchmark designed to develop and evaluate LLM-powered personalized assistants. Unlike existing datasets for personalization research, HiCUPID satisfies the five desiderata of an LLM-backed personalized assistant, offering new opportunities for advancing LLM personalization. HiCUPID additionally provides a Llama-3.2-based automated evaluation model whose assessment is well-aligned human preferences. Extensive experiments with HiCUPID reveal shortcomings and potentials of current state-of-the-art LLMs and common approaches to personalization.

Limitations and Potential Risks

Many of human evaluators noted that general responses could be preferred by humans over over-personalized responses. The matter of how much personalization a typical human usual prefers when interacting with chatbots and assistants is a complex sociological question and remains an open research topic that is beyond the scope of our paper. We observed that when GPT-4o is prompted to generate a personalized answer based on a user's persona that contains negative sentiments (*e.g.,* dislikes, does not enjoy, does not have, is not interested in, etc.), the generated answer sometimes omitted this person, instead of explicitly stating the user's negative stance. This is likely due to the widely-known struggle of LLMs to comprehend and generate negations. The subsequent versions of HiCUPID will be augmented to include more personalized answers with negative sentiments. Lastly, the failure of DPO on the majority of LLMs highlights the difficulty of training LLMs with reinforcement learning (RL)-based approaches. In the future, we aim to develop a reward model specifically for HiCUPID, such that it can be used for RL-based training.

If more advanced personalized assistants become available, it might become easier to extract personal identifiable information (PII) from personalized LLMs. While HiCUPID, being a synthetic dataset, does not contain any PII of real human users, improving the degree of personalization will inevitably aggravate data privacy concerns. To prevent privacy risks and potential misuse of PII, the development of personalized assistants must be accompanied by research on privacy-preserving measures, such as differential privacy, homomorphic encryption, or data anonymization. Moreover, the inherent bias in AI design or implementation may be amplified during the personalization process. When generating a personalized response, the assistant will rely on its prior knowledge of what a human user with specific personality traits may prefer. In doing so, the assistant could precipitate biased information or stereotypes regarding personas, and thus, additionally grounding responses to be ethical and fair via instruction tuning is necessary. Additionally, we tried to make our users as diverse as possible, but there may exist demographic groups that are potentially underrepresented in our dataset. To dynamically generate new data to compensate for underrepresented demographics, we provided all of the prompt templates and necessary resources. Lastly, over-reliance on personalized assistants may harm the critical thinking and decision making ability of human users.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (Ministry of Science and ICT, MSIT) (No.2022R1A3B1077720, 2022R1A5A708390811), Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (Ministry of Science and ICT, MSIT) (No.RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University), RS-2022-II220959, RS-2025-02263754, Human-Centric Embodied AI Agents with Autonomous Decision-Making) the BK21 FOUR program of the Education and Research Program for Future ICT Pioneers, Seoul National University, Samsung Electronics Co., Ltd (IO240311-09242-01), a grant from the Yang Young Foundation.

References

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

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6.

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

Fabio Crestani, Mounia Lalmas, Cornelis J Van Rijsbergen, and Iain Campbell. 1998. "is this document relevant?… probably" a survey of probabilistic models in information retrieval. *ACM Computing Surveys (CSUR)*, 30(4):528–552.

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

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

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jinlan Fu, See Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. Gptscore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576.

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

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

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Pegah Jandaghi, XiangHai Sheng, Xinyi Bai, Jay Pujara, and Hakim Sidahmed. 2023. Faithful persona-based conversational dataset generation with large language models. *arXiv preprint arXiv:2312.10007*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

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

Sudipta Kar, Gustavo Aguilar, Mirella Lapata, and Thamar Solorio. 2020. Multi-view story characterization from movie plot synopses and reviews. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5629–5646.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

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

Junyi Li, Ninareh Mehrabi, Charith Peris, Palash Goyal, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta. 2023b. On the steerability of large language models toward data-driven personas. *arXiv preprint arXiv:2311.04978*.

Zhiyu Li, Yanfang Chen, Xuan Zhang, and Xun Liang. 2023c. Bookgpt: A general framework for book recommendation empowered by large language model. *Electronics*, 12(22):4654.

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

Qijiong Liu, Nuo Chen, Tetsuya Sakai, and Xiao-Ming Wu. 2024. 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*, pages 452–461.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68.

Shengyu Mao, Ningyu Zhang, Xiaohan Wang, Mengru Wang, Yunzhi Yao, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. Editing personality for large language models.

Behrad Moniri, Hamed Hassani, and Edgar Dobriban. 2024. Evaluating the performance of large language models via debates. *arXiv preprint arXiv:2406.11044*.

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

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

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

Hongjin Qian, Xiaohe Li, Hanxun Zhong, Yu Guo, Yueyuan Ma, Yutao Zhu, Zhanliang Liu, Zhicheng Dou, and Ji-Rong Wen. 2021. Pchatbot: a large-scale dataset for personalized chatbot. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 2470–2477.

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

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

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

Alireza Salemi, Surya Kallumadi, and Hamed Zamani. 2024a. Optimization methods for personalizing large language models through retrieval augmentation. *Preprint*, arXiv:2404.05970.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024b. LaMP: When large language models meet personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7370–7392, Bangkok, Thailand. Association for Computational Linguistics.

Alireza Salemi and Hamed Zamani. 2024. Comparing retrieval-augmentation and parameter-efficient fine-tuning for privacy-preserving personalization of large language models. *Preprint*, arXiv:2409.09510.

Weiyan Shi, Yu Li, Saurav Sahay, and Zhou Yu. 2021. Refine and imitate: Reducing repetition and inconsistency in persuasion dialogues via reinforcement learning and human demonstration. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3478–3492.

Zhaoxuan Tan and Meng Jiang. 2023. User modeling in the era of large language models: Current research and future directions. *arXiv preprint arXiv:2312.11518*.

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

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

Danqing Wang, Kevin Yang, Hanlin Zhu, Xiaomeng Yang, Andrew Cohen, Lei Li, and Yuandong Tian. 2023. Learning personalized story evaluation. *arXiv preprint arXiv:2310.03304*.

T Wolf. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

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

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

Kevin Yang, Dan Klein, Nanyun Peng, and Yuandong Tian. 2023. Doc: Improving long story coherence with detailed outline control. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3378–3465.

Runzhe Yang, Jingxiao Chen, and Karthik Narasimhan. 2021. Improving dialog systems for negotiation with personality modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 681–693.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.

Kai Zhang, Yangyang Kang, Fubang Zhao, and Xiaozhong Liu. 2024. Llm-based medical assistant personalization with short-and long-term memory coordination. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2386–2398.

Saizheng Zhang. 2018. Personalizing dialogue agents: I have a dog, do you have pets too. *arXiv preprint arXiv:1801.07243*.

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

Feng Zhu, Yan Wang, Chaochao Chen, Guanfeng Liu, and Xiaolin Zheng. 2020. A graphical and attentional framework for dual-target cross-domain recommendation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3001–3008.

Appendix

A1 Extended Related Works

This section, an extension of Section 2 in the main paper, discusses previous approaches in personalization research in more detail. In the domain of personalized recommendation, the rating history of a user is provided within the prompt in the form of few-shot instances (Dai et al., 2023; Kang et al., 2023). PERSE (Wang et al., 2023) and BookGPT (Li et al., 2023c) leverage past book reviews to perform personalized store evaluations and book recommendations, respectively. Because the prompting-based approaches are inherently limited by the context length of LLMs, RAG-based methods were devised to address this shortcoming. PEARL (Mysore et al., 2023) calibrates the retriever to select past documents authored by the user. LaMP and its subsequent works (Salemi et al., 2024b; Salemi and Zamani, 2024; Salemi et al., 2024a) show that RAG is a promising avenue toward personalized generation that can also guarantee the privacy of user's personal information. The family of profile-augmentation techniques (Richardson et al., 2023; Liu et al., 2024) further improves RAG-based approaches by supplying the LLM with a summary of user persona and interactions.

OPPU (Tan et al., 2024) trains one LoRA (Hu et al.) module per user to store user-specific information and integrates this parametric knowledge base with non-parametric knowledge from a retriever. Similarly, PEFT-RAG (Salemi and Zamani, 2024) first fine-tunes the LLM with the LoRA adapter on user-specific information and then applies RAG. Li *et al.* (Li et al., 2023b) and Tang *et al.* (Tang et al., 2023) utilize data-driven methods to extract persona information more compactly to reduce noisy and fine-grained learning signals.

A2 Persona Dimensions of HiCUPID

Following is the list of persona dimensions used to define synthetic users in HiCUPID. The personas within each persona dimension are provided in a .xlsx file in Supplementary Materials.

1. Sports
2. Fashion
3. Electronics
4. Game
5. Movie
6. Major
7. Fitness
8. Art
9. Music
10. Politics
11. Beauty
12. Animal
13. Environment
14. Religion
15. Family
16. Self-improvement
17. Travel
18. Car
19. Technology
20. Book
21. Social Media
22. Cooking & Baking
23. Food
24. Forms of Living
25. Finance

A3 How HiCUPID Tests 5 Desiderata of Personalized Assistant

Below, we detail how our dataset configuration and evaluation criteria in the evaluation prompt (Figure A15) together probe the five desiderata of a personalized assistant, outlined in Section 3.

(a) *Adherence to User Information (AUI)*: This desideratum is in fact provided as a part of the evaluation prompt: 1. Personalization: Does the response effectively consider the user's provided personal information? Therefore, the LLM must adhere to the user's personal information for its response to meet this "Personalization" criterion in the evaluation prompt.

(b) *Understanding of Implicit Information (UII)*: The dialogue history contains implicit cues to the user's personal information, instead of explicit and structured personal information. Therefore, in order to meet the "Personalization" criterion in the evaluation prompt, the LLM must correctly understand the personal information implicitly embedded in the dialogue history.

(c) *Reasoning from Multiple Information (MI)*: multi-info QA pairs include questions that require simultaneously considering a user's persona and profile to be answered properly. Therefore, if the LLM's response on the multi-info QA pairs meets the "Personalization" criterion, we can deduce that LLM picked up on and reasoned from two pieces of persona information.

(d) *Long-context Modeling Capacity (LC)*: Personal information appears scattered throughout the dialogue history that contains, on average, 15k tokens, and thus, to satisfy the "Personalization" criterion, the LLM must be able to model long contextual information when generating responses.
(e) *Proactiveness of Responses (PR)*: All QA pairs in HiCUPID are designed such that the LLM must provide proactive answers to the user's question. Therefore, satisfying the "personalization" and "logical validity" criteria in the evaluation prompt can be interpreted as the generated responses being proactive and logically sound.

A4 Prompt Templates for HiCUPID Generation

Figures A5, A6, and A7 are the prompt templates used to generate dialogues in HiCUPID with GPT-4o. Figures A8 and A9 are the prompt templates used to generate single-info persona and schedule QA pairs. Figure A4 is the prompt template used to generate logical and realistic persona-profile combinations, and Figure A10 is used to generate multi-info QA pairs. The user's persona is included under [User's Characteristics], and the user's profile and schedule are included under [User's Profile] and [User's Schedule], respectively.

A5 Human and GPT-4o Evaluation Protocols

Human Evaluation: The human evaluators in our study are comprised of undergraduate and graduate-level students. The evaluators were notified that the results would be included in an academic research paper, but the topic of the research paper was not revealed to them. They were volunteers without explicit payments. Please refer to Figure A14 for the survey template used for Human Evaluation.
GPT-4o & Distilled Llama-3.2 Evaluation Protocol: The prompt template for GPT-4o and Llama-3.2 evaluation is also provided in Figure A15.

A6 Proxy Evaluation Model Training Protocol

A total of 400k GPT-4o evaluation samples are used to train the Llama-3.2-3B-based proxy evaluation model. The samples are from the inference results in the following experimental settings:
• GPT-4o-mini: Zero- and few-shot inference with prompt in the user role.
• Llama, Mistral, Qwen: Zero- and few-shot inference with prompt in the user role; BM25- and Contriever-based utterance-level retrieval with $k = 5$; SFT (LR=1e-4) and DPO (LR=1e-5) with LoRA$_r = 256$.
The detailed hyperparameter settings for training evaluation model are summarized in Table A1.

A7 Experimental Settings and Hyperparameters

• **Zero-shot:** We prompt the base LLM to generate a personalized response based on the entire dialogue history provided as a part of the prompt. The prompt for zero-shot inference is presented in Figures A11. The prompt with the instruction in the user role is used as the default prompt to obtain the main results.
• **Few-shot:** We supply the base LLM with $n \in \{1, 3, 5\}$ sets of QA pair examples as in-context demonstrations to assist in generating personalized responses. Each set includes a question, a personalized answer, and a general answer for each QA pair type—persona QA, schedule QA, and multi-info (persona+profile) QA—to encompass all QA pair types in HiCUPID. The example QA pairs are randomly sampled from the train set and fixed across all experiments. The prompt for a few-shot inference is presented in Figures A12 and A13. The prompt with the instruction in the user role (Figure A12) is used as the default prompt to obtain the main results.
• **BM25** (Crestani et al., 1998): We select top-k question-relevant messages from the entire dialogue history based on a rule-based ranking algorithm. The chosen messages are then used in place of the dialogue history. An individual dialogue and an individual utterance are used as a unit of retrieval. For dialogue-level retrieval, we test three different numbers of retrieved units: $k \in \{1, 3, 5\}$. For utterance-level retrieval, we experiment with $k \in \{5, 15, 25\}$. The main results are reported using utterance-level retrieval with $k = 5$.
• **Contriever** (Izacard et al., 2021): We select top-k question-relevant messages from the entire dialogue history based on a model-based cosine similarity measure. The selected messages are then fed into the base LLM in the same way as in BM25. We investigate the same set of retrieval settings as those used for BM25. The main results are reported using utterance-level retrieval with $k = 5$.

Hyperparameter	SFT	DPO	SFT+DPO	Proxy Evaluation Model
Batch Size	256	256	256	1024
Train Epochs	1	1	1	1
Optimizer	AdamW	AdamW	AdamW	AdamW
LR	{1e-6, 3e-6, 1e-5, 3e-5, **1e-4**, 3e-4}	{**1e-6**, 3e-6, 1e-5, 3e-5, 1e-4, 3e-4}	{1e-6, **3e-6**, 1e-5, 3e-5, 1e-4, 3e-4}	1e-4
Adam β_1	0.9	0.9	0.9	0.9
Adam β_2	0.999	0.999	0.999	0.999
Weight Decay	0.0	0.0	0.0	0.0
LR Scheduler	cosine annealing	cosine annealing	cosine annealing	cosine annealing
Warmup Ratio	0.03	0.03	0.03	0.03
Max Grad Norm	0.3	0.3	0.3	0.3
LoRA_r (Rank)	{8, 16, 32, 64, 128, **256**}	{8, 16, 32, 64, **128**, 256}	{8, 16, 32, 64, 128, **256**}	128
LoRA_α (Alpha)	$\text{LoRA}_r \times 2$	$\text{LoRA}_r \times 2$	$\text{LoRA}_r \times 2$	$\text{LoRA}_r \times 2$
LoRA Target Modules	all-linear	all-linear	all-linear	all-linear
LoRA Dropout	0.05	0.05	0.05	0.05

Table A1: Hyperparameter settings for fine-tuning open-source LLMs on HiCUPID and for training the Llama-3.2-3B proxy evaluation model.

- **Supervised Fine-Tuning (SFT)**: We fine-tune the base LLM with the entire dialogue history and the user's question as the input x and the ground-truth personalized answer as the output y. The question and answer pairs are from the train split of HiCUPID. We adopt LoRA (Hu et al.), a popular PEFT module used to fine-tune LLMs on downstream tasks. The detailed hyperparameter settings for SFT are summarized in Table A1. We conducted an experiment to analyze the effect of the number of training epochs, but the performance change was minimal. The main results are obtained with LR=1e-4 and $\text{LoRA}_r = 256$, which is the best hyperparameter setting searched on the Llama model.
- **Direct Preference Optimization (DPO)** (Rafailov et al., 2024): We fine-tune the base LLM with the modified dialogue history and the user's question as the input x, the personalized answer as the output y_{chosen}, and the general answer as the output y_{rejected}. For the modified dialogue history, we excluded 15 question-irrelevant persona dialogues from the entire dialogue history due to the GPU VRAM limitations. We also adopt LoRA for training DPO. The detailed hyperparameter settings for DPO are summarized in Table A1. We conducted an experiment to analyze the effect of the number of training epochs, but the performance change was minimal. The main results are obtained with LR=1e-6 and $\text{LoRA}_r = 128$, which is the best hyperparameter setting searched on the Llama model.
- **SFT+DPO** (Rafailov et al., 2024): In the previous step, we obtained a LoRA SFT model trained on personalized answers as the ground truth. The LoRA adapters are merged into the base LLM, and the merged model undergoes additional preference fine-tuning using DPO. For DPO, we fine-tune the merged model with the modified dialogue history and the user's question as the input x, the personalized answer as the output y_{chosen}, and the general answer as the output y_{rejected}. For the modified dialogue history, we excluded 15 question-irrelevant persona dialogues from the entire dialogue history due to the GPU VRAM limitations. We also adopt LoRA for SFT+DPO. The detailed hyperparameter settings for SFT+DPO are summarized in Table A1. The main results are obtained with LR=3e-6 and $\text{LoRA}_r = 256$, which is the best hyperparameter setting searched on the Llama model.

With the exception of RAG-based methods (BM25, Contriever), the entire dialogue history is provided as a part of the prompt at inference time.

A8 Extended GPT-4o and Llama-3.2 Evaluation Results

Due to the page limit, we only reported S_{GPT} and S_{Llama} in the main paper. We report Model Win, Tie, and Model Lose (GT Win) rates separately for Test Sets 1 and 2 in Tables A3 and A4, respectively.

A9 Retrieval Performance

The performance of each retrieval setting is reported in Table A2.

A10 Extended Ablation Results

Due to the page limit, the ablation study results only on Test Set 1 were reported in the main paper. Here, the full results of ablation studies on Test Sets 1 and 2 are reported. Table A5 reports the full results of replacing the whole dialogue history with the gold dialogue that is relevant to each QA pair. Table A6 compares the full results of placing the zero- and few-shot prompt in the user role against those obtained by placing the prompt under the system prompt. The results in these tables clearly show that LLMs' struggle with long context makes it difficult to model the whole dialogue history or follow the instruction provided before the whole dialogue history.

Table A7 reports the full result of changing the number of demonstrations in the few-shot prompt. Table A8 includes the full result of changing the retrieval setting. Lastly, Tables A9, A10, and A11 report the full results of extensive hyperparameter search on SFT, DPO, and SFT+DPO, respectively.

A11 Computational Environment and Cost of Research

Our experiments were conducted on NVIDIA H100, L40, and A40 GPUs. With the exception of GPT variants, all models were downloaded from the Huggingface library. The GPT variants (GPT-4o and GPT-4o-mini) were accessed via the OpenAI API. The total cost of dataset generation and evaluation with GPT-4o was approximately US$1,500.

A12 License

Llama-3.1&3.2 models are licensed under the Llama-3.1&3.2 COMMUNITY LICENSE. Mistral and Qwen are licensed under Apache 2.0. GPT-4o and GPT-4o-mini are licensed under OpenAI.

A13 Use of AI Writing Assistant

The use of ChatGPT-4o was limited to sentence-level paraphrasing and word-level synonym search. No additional AI assistant was involved in research, coding, or writing.

Unit	#	BM25					Contriever				
		Persona	Schedule	Multi-Info		Total	Persona	Schedule	Multi-Info		Total
				Persona	Profile				Persona	Profile	
Dialogue	1	42.4	93.3	34.5	8.6	52.5	64.3	40.3	42.8	22.1	54.3
Dialogue	3	58.0	99.7	49.2	19.4	65.4	80.6	79.9	59.7	45.0	76.9
Dialogue	5	65.4	100.0	57.1	27.1	71.1	86.0	94.1	66.8	59.6	85.2
Utterance	5	50.0	99.6	38.0	23.0	60.0	82.1	91.2	65.4	29.1	80.0
Utterance	15	72.7	100.0	63.8	46.4	77.3	92.6	100.0	79.2	48.4	90.8
Utterance	25	83.6	100.0	78.8	63.9	86.1	96.0	100.0	85.0	59.4	94.0

Table A2: Retrieval performance under various retrieval settings. The performance is measured through the percentage of retrieval results where the associated dialogues are retrieved. For multi-info QA pairs with two associated dialogues, one for the persona and the other for the profile, the scores for persona-hit and profile-hit are reported separately. For the profile, we consider it retrieved as long as any subset of the five profiles is extracted. The total score is computed as a weighted average as follows: persona $\times \frac{25}{40}$ + schedule $\times \frac{10}{40}$ + (multi-info persona + multi-info profile) / 2 $\times \frac{5}{40}$.

Model	Setting	GPT-4o Score (S_{GPT})										LLaMA Score (S_{Llama})									
		Persona				Schedule	Multi-Info				Total	Persona				Schedule	Multi-Info				Total
		Score	Win	Tie	Lose	Score	Score	Win	Tie	Lose	Score	Score	Win	Tie	Lose	Score	Score	Win	Tie	Lose	Score
GPT-4o-mini	0-shot	42.1	29.3	25.5	45.1	9.5	4.4	1.9	5.0	93.1	28	44.7	27.4	34.6	38.0	8.8	10.8	5.6	10.4	84.0	30.4
	few-shot	40.5	27.8	25.5	46.8	76.1	4.2	2.2	4.0	93.8	35.3	42.6	24.9	35.4	39.7	75.4	11.4	5.7	11.4	83.0	37.5
Llama-3.1-8B	0-shot	38.0	32.4	11.2	56.4	13.9	3.5	1.9	3.1	95.0	25.9	39.7	30.8	17.7	51.5	9.4	8.1	5.4	5.5	89.1	27.0
	few-shot	39.4	34.0	10.7	55.3	49.8	6.3	4.4	3.8	91.8	31.6	38.8	29.4	18.8	51.8	48.3	12.3	8.2	8.2	83.6	31.8
	BM25	29.7	19.5	20.3	60.2	84.3	2.3	0.7	3.1	96.2	29.4	34.1	18.1	32.1	49.9	78.5	6.1	2.2	7.8	90.0	31.9
	Contriever	38.8	27.8	21.9	50.3	75.4	4.2	2.3	3.8	93.9	34.2	42.6	26.2	32.8	41.0	70.3	9.8	5.0	9.6	85.4	36.6
	SFT	36.2	26.8	18.8	54.4	88.0	12.4	10.2	4.3	85.4	35.2	36.5	26.8	19.5	53.7	87.5	15.8	13.2	5.1	81.7	35.7
	DPO	24.8	9.4	30.7	59.9	4.9	2.1	0.2	3.8	96.0	16.4	34.8	10.4	48.9	40.8	4.2	6.4	0.5	11.8	87.8	23.1
	SFT+DPO	49.1	37.4	23.4	39.2	98.6	14.5	12.2	4.6	83.2	44.8	48.1	35.0	26.1	38.9	98.1	18.4	16.1	4.6	79.4	44.6
Mistral-7B	0-shot	20.9	8.8	24.3	66.9	0.0	1.5	0.4	2.2	97.4	13.3	30.5	8.9	43.2	48.0	0.0	3.8	0.4	6.8	92.8	19.5
	few-shot	28.6	12.4	32.3	55.3	6.3	3.5	1.1	4.8	94.1	19.1	36.2	11.2	50.0	38.8	5.6	7.6	0.9	13.4	85.8	24.2
	BM25	41.0	31.6	18.9	49.5	8.6	4.9	2.5	4.8	92.7	27.3	43.7	29.1	29.1	41.7	6.3	9.0	3.6	10.8	85.6	29.2
	Contriever	48.8	39.2	19.3	41.5	7.9	7.4	4.0	6.9	89.1	32.4	51.6	38.0	27.1	34.8	5.9	13.6	6.9	13.5	79.6	34.7
	SFT	27.6	16.3	22.5	61.2	99.8	15.1	12.5	5.3	82.2	31.6	31.2	17.8	26.7	55.4	99.8	19.7	16.6	6.3	77.1	34.4
	DPO	8.2	6.9	2.6	90.5	2.2	0.2	0.2	0.0	99.8	5.4	6.4	5.2	2.3	92.5	1.4	0.5	0.3	0.3	99.4	4.2
	SFT+DPO	44.7	31.9	25.6	42.5	99.7	17.6	15.0	5.2	79.8	42.6	44.8	30.3	29.1	40.7	99.8	20.4	17.3	6.3	76.4	43.0
Qwen-2.5-7B	0-shot	26.6	11.9	29.5	58.6	0.0	3.0	0.8	4.5	94.7	17	34.6	11.0	47.3	41.7	0.0	6.1	1.0	10.2	88.8	22.4
	few-shot	24.6	10.4	28.4	61.2	29.8	2.1	0.3	3.6	96.1	19.4	32.6	9.3	46.7	44	28.7	4.8	0.6	8.4	91.0	24.6
	BM25	30.6	13.4	34.5	52.1	0.4	3.0	0.3	5.3	94.4	19.6	37.7	11.4	52.6	36	0.1	6.8	0.3	12.9	86.8	24.4
	Contriever	33.6	16.4	34.5	49.1	0.2	3.4	0.6	5.6	93.8	21.5	39.6	14.3	50.6	35.1	0.1	7.4	0.6	13.4	85.9	25.7
	SFT	35.7	24.8	21.8	53.4	99.7	25.4	21.8	7.2	71.0	37.9	38.3	26	24.6	49.4	99.8	33.3	29.2	8.2	62.6	40.6
	DPO	36.6	15.0	43.2	41.8	0.0	8.8	0.7	16.1	83.2	24.0	38.0	11.6	52.9	35.5	0.0	12.4	0.2	24.4	75.4	25.3
	SFT+DPO	43.1	32.8	20.5	46.7	99.8	34.0	29.9	8.2	61.9	43.6	43.2	31.6	23.0	45.3	99.9	38.1	34.3	7.5	58.2	44.2

Table A3: Extended evaluation results on **Test Set 1 (Seen User/Unseen QA)**.

Model	Setting	GPT-4o Score (S_{GPT})										LLaMA Score (S_{Llama})									
		Persona				Schedule	Multi-Info				Total	Persona				Schedule	Multi-Info				Total
		Score	Win	Tie	Lose	Score	Score	Win	Tie	Lose	Score	Score	Win	Tie	Lose	Score	Score	Win	Tie	Lose	Score
GPT-4o-mini	0-shot	42.0	29.2	25.8	45.1	9.2	4.9	1.9	6.0	92.1	28.0	45.5	26.9	37.1	36.0	8.8	11.9	5.5	12.7	81.8	31.0
	few-shot	40.7	27.4	26.6	46.0	76.2	5.1	2.0	6.2	91.8	35.6	43.7	25.1	37.1	37.8	75.2	11.6	5.4	12.4	82.2	38.1
Llama-3.1-8B	0-shot	38.6	32.9	11.2	55.8	13.9	3.2	1.1	4.2	94.7	26.2	40.6	32	17.1	50.9	9.8	8.7	5.3	6.9	87.8	27.7
	few-shot	39.3	33.5	11.6	54.9	49.2	7.5	5.3	4.5	90.2	31.6	39.8	30.8	17.9	51.2	47.8	14.1	9.6	9.0	81.4	32.6
	BM25	29.9	19.5	20.9	59.6	84.0	2.7	1.0	3.5	95.5	29.6	35.0	18.4	33.1	48.5	78.0	7.0	2.1	9.8	88.1	32.5
	Contriever	37.9	27.2	21.4	51.4	77.4	4.5	2.1	4.8	93.1	33.9	43.1	26.6	32.9	40.4	71.8	9.1	4.5	9.2	86.3	37.1
	SFT	34.4	25.9	17.0	57.1	88.4	12	10.2	3.7	86.2	34	34.8	26.3	17.0	56.7	87.5	14.4	12.4	4.0	83.6	34.5
	DPO	25.1	9.6	30.9	59.5	5.8	2.4	0.2	4.4	95.4	16.7	35.1	10.3	49.6	40.1	5.1	6.3	0.9	10.8	88.3	23.4
	SFT+DPO	47.1	36.1	22.0	42.0	98.7	18.6	16.0	5.1	78.9	44.1	47.0	35.2	23.6	41.2	98.1	22.4	19.4	6.0	74.6	44.4
Mistral-7B	0-shot	21.8	9.6	24.3	66.1	0.0	1.8	0.3	3.0	96.6	13.8	30.8	9.6	42.5	48	0.0	5.0	0.5	9.0	90.5	19.9
	few-shot	28.9	13	31.6	55.3	8.0	3.8	0.7	6.2	93.1	19.5	36.6	11.0	51.2	37.8	6.6	8.2	1.1	14.2	84.7	24.7
	BM25	40.6	30.5	20.1	49.4	8.2	5.9	2.8	6.2	91.0	27.1	43.6	29	29.1	41.9	5.9	10.4	4.5	11.8	83.8	29.3
	Contriever	48.6	38.8	19.5	41.7	8.6	8.4	5.4	6.0	88.6	32.5	50.9	37.2	27.5	35.3	6.4	15.1	8.6	13.1	78.3	34.5
	SFT	27.6	17.4	20.2	62.3	99.9	13.3	11.7	3.2	85.1	31.4	31.5	19.8	23.3	56.9	100	18.1	15.4	5.4	79.2	34.4
	DPO	8.1	6.8	2.6	90.6	2.0	0.3	0.2	0.1	99.7	5.3	6.5	5.2	2.4	92.3	1.4	0.6	0.6	0.2	99.3	4.3
	SFT+DPO	43.2	31.1	24.2	44.7	99.9	17.8	15.0	5.6	79.4	41.7	43.6	30.2	26.8	43.0	99.9	22.5	19.5	5.9	74.6	42.5
Qwen-2.5-7B	0-shot	27.8	13.1	29.4	57.5	0.0	2.5	0.5	4.1	95.4	17.7	34.9	11.6	46.7	41.7	0.0	6.4	1.0	11.0	88.1	22.6
	few-shot	25.1	10.9	28.4	60.7	27.3	2.1	0.5	3.3	96.2	19.4	33.4	9.8	47.1	43.1	25.4	6.2	0.7	10.9	88.4	24.8
	BM25	31.1	13.6	35.1	51.3	0.4	3.1	0.5	5.2	94.3	19.9	38.1	11.8	52.6	35.6	0.2	7.8	0.4	14.7	84.9	24.8
	Contriever	34.0	17.0	34.1	48.9	0.4	4.0	0.6	6.7	92.6	21.8	40.8	15.1	51.3	33.6	0.1	8.1	0.8	14.6	84.6	26.5
	SFT	34.2	23.9	20.8	55.4	99.9	24.9	21.6	6.6	71.8	37	38.3	26.8	22.8	50.3	99.9	30.8	27.5	6.6	65.8	40.3
	DPO	37.0	14.9	44.1	40.9	0.2	8.8	0.6	16.4	83.0	24.3	39.0	12.6	52.8	34.6	0.1	12.6	0.2	24.8	75.0	26.0
	SFT+DPO	41.9	32.5	18.8	48.8	99.8	33.9	31.3	5.3	63.4	42.9	42.9	32.4	21.0	46.6	99.8	37.8	34.6	6.6	58.9	44.0

Table A4: Extended evaluation results on **Test Set 2 (Unseen User/Unseen QA)**.

Model	Shot	Type	Test Set 1 (Seen User/Unseen QA) Persona	Schedule	Multi-Info	Total	Test Set 2 (Unseen User/Unseen QA) Persona	Schedule	Multi-Info	Total
GPT-4o-mini	0-shot	Gold	68.0	95.4	26.1	57.6	68.9	94.9	26.8	58.2
	0-shot	Whole	44.7	8.8	10.8	30.4	45.5	8.8	11.9	31.0
	3-shot	Gold	65.5	100.0	22.8	56.3	66.5	100.0	23.2	56.9
	3-shot	Whole	42.6	75.4	11.4	37.5	43.7	75.2	11.6	38.1
Llama-3.1-8B	0-shot	Gold	61.6	77.3	16.7	50.3	62.5	77.3	17.6	50.9
	0-shot	Whole	39.7	9.4	8.1	27.0	40.6	9.8	8.7	27.7
	3-shot	Gold	61.8	99.8	23.2	54.0	62.7	99.6	23.7	54.6
	3-shot	Whole	38.8	48.3	12.3	31.8	39.8	47.8	14.1	32.6
Mistral-7B	0-shot	Gold	46.2	7.8	8.8	30.9	47.0	7.9	10.1	31.7
	0-shot	Whole	30.5	0.0	3.8	19.5	30.8	0.0	5.0	19.9
	3-shot	Gold	44.0	77.6	9.9	38.4	44.5	75.5	11.0	38.6
	3-shot	Whole	36.2	5.6	7.6	24.2	36.6	6.6	8.2	24.7
Qwen-2.5-7B	0-shot	Gold	52.1	1.0	15.9	34.7	52.8	0.9	16.3	35.1
	0-shot	Whole	34.6	0.0	6.1	22.4	34.9	0.0	6.4	22.6
	3-shot	Gold	49.7	6.4	12.4	33.4	50.2	6.1	13.2	33.8
	3-shot	Whole	32.6	28.7	4.8	24.6	33.4	25.4	6.2	24.8

Table A5: Full ablation results on LLMs' long-context handling ability.

Model	Shots	Type	Test Set 1 (Seen User/Unseen QA) Persona	Schedule	Multi-Info	Total	Test Set 2 (Unseen User/Unseen QA) Persona	Schedule	Multi-Info	Total
Llama-3.1-8B	0-shot	User	30.2	1.7	3.8	19.5	31.4	1.9	4.9	20.5
	0-shot	System	39.7	9.4	8.1	27.0	40.6	9.8	8.7	27.7
	3-shot	User	30.6	2.7	4.2	20.0	30.8	2.5	4.7	20.1
	3-shot	System	38.8	48.3	12.3	31.8	39.8	47.8	14.1	32.6
Mistral-7B	0-shot	User	34.5	0.0	6.0	22.3	34.6	0.0	6.8	22.5
	0-shot	System	30.5	0.0	3.8	19.5	30.8	0.0	5.0	19.9
	3-shot	User	37.3	12.6	11.5	26.3	38	12.7	10.9	26.7
	3-shot	System	36.2	5.6	7.6	24.2	36.6	6.6	8.2	24.7
Qwen-2.5-7B	0-shot	User	33.8	0.0	8.0	22.1	34.3	0.0	9.3	22.6
	0-shot	System	34.6	0.0	6.1	22.4	34.9	0.0	6.4	22.6
	3-shot	User	33.8	0.1	7.9	22.1	34.7	0.1	8.4	22.8
	3-shot	System	32.6	28.7	4.8	24.6	33.4	25.4	6.2	24.8

Table A6: Ablation on prompt roles in instructions and few-shot examples.

Model	Shots	Test Set 1 (Seen User/Unseen QA) Persona	Schedule	Multi-Info	Total	Test Set 2 (Unseen User/Unseen QA) Persona	Schedule	Multi-Info	Total
GPT-4o-mini	0-shot	44.7	8.8	10.8	30.4	45.5	8.8	11.9	31.0
	1-shot	43.1	78.9	10.4	38.1	43.2	78.2	10.6	38.1
	3-shot	42.6	75.4	11.4	37.5	43.7	75.2	11.6	38.1
	5-shot	43.4	65.6	10.3	36.6	43.9	66.1	11.9	37.2
Llama-3.1-8B	0-shot	39.7	9.4	8.1	27	40.6	9.8	8.7	27.7
	1-shot	37.3	45.3	9.9	30.2	37.7	44.7	9.9	30.4
	3-shot	38.8	48.3	12.3	31.8	39.8	47.8	14.1	32.6
	5-shot	33.4	34.7	11.6	26.6	34.8	34.7	12.2	27.6
Mistral-7B	0-shot	30.5	0.0	3.8	19.5	30.8	0.0	5.0	19.9
	1-shot	33.2	0.4	5.3	21.4	34.5	0.4	6.0	22.4
	3-shot	36.2	5.6	7.6	24.2	36.6	6.6	8.2	24.7
	5-shot	36.6	7.0	8.2	24.8	37.2	7.5	8.8	25.3
Qwen-2.5-7B	0-shot	34.6	0.0	6.1	22.4	34.9	0.0	6.4	22.6
	1-shot	34.0	0.0	5.3	21.9	35.0	0.1	6.2	22.7
	3-shot	32.6	28.7	4.8	24.6	33.4	25.4	6.2	24.8
	5-shot	32.4	21.0	5.1	23.5	33.5	19.6	5.1	24.0

Table A7: Full ablation results of zero- and few-shot inference.

Method	Unit	Number	Test Set 1 (Seen User/Unseen QA) Persona	Schedule	Multi-Info	Total	Test Set 2 (Unseen User/Unseen QA) Persona	Schedule	Multi-Info	Total
BM25	Dialogue	1	30.9	85.3	6.4	30.8	32.1	86.4	7.5	31.8
	Dialogue	3	32.7	83.8	4.8	31.5	33.2	81.8	5.2	31.6
	Dialogue	5	30.8	78.4	4.6	29.6	32.3	79.1	4.7	30.7
	Utterance	5	34.1	78.5	6.1	31.9	35.0	78.0	7.0	32.5
	Utterance	15	30.7	61.9	4.2	27.5	31.7	62.1	5.4	28.2
	Utterance	25	30.3	38.0	4.2	24.2	30.4	38.7	4.6	24.4
Contriever	Dialogue	1	40.0	38.0	11.2	31.1	40.1	36.8	11.2	31.0
	Dialogue	3	32.2	66.9	7.2	29.4	33.6	66.0	6.3	30.0
	Dialogue	5	30.6	74.0	6.0	29.2	31.9	73.8	5.3	29.9
	Utterance	5	42.6	70.3	9.8	36.6	43.1	71.8	9.1	37.1
	Utterance	15	33.1	59.4	4.8	28.7	33.5	60.3	5.3	29.1
	Utterance	25	30.8	34.5	4.3	24.1	31.3	35.8	5.0	24.7

Table A8: Full ablation results of various settings of Retrieval-augmented Generation.

Hyperparam.	Value	Test Set 1 (Seen User/Unseen QA) Persona	Schedule	Multi-Info	Total	Test Set 2 (Unseen User/Unseen QA) Persona	Schedule	Multi-Info	Total
Learning Rate	1e-6	16.1	19.6	3.4	12.9	16.9	18.2	3.8	13.3
	3e-6	22.6	65.9	5.0	23.0	22.6	67.3	5.1	23.2
	1e-5	30.0	67.0	10.1	28.4	29.3	66.0	10.0	27.8
	3e-5	33.2	84.4	12.3	32.9	34.1	85.0	10.8	33.3
	1e-4	36.5	87.5	15.8	35.7	34.8	87.5	14.4	34.5
	3e-4	34.3	88.3	12.9	34.1	32.9	88.6	13.8	33.3
Rank	8	31.7	70.6	10.8	30.0	31.1	68.9	9.9	29.3
	16	32.8	69.1	11.4	30.5	32.3	69.8	11.3	30.4
	32	33.3	82.6	12.4	32.7	33.8	81.6	11.2	32.7
	64	34.0	85.0	13.4	33.6	34.7	84.4	13.0	33.9
	128	36.7	88.4	16.0	36.0	35.3	88.2	13.5	34.8
	256	36.5	87.5	15.8	35.7	34.8	87.5	14.4	34.5

Table A9: Ablation on hyperparameters in training Llama-3.1 model with LoRA-based Supervised Fine-tuning.

Hyperparam.	Value	Test Set 1 (Seen User/Unseen QA) Persona	Schedule	Multi-Info	Total	Test Set 2 (Unseen User/Unseen QA) Persona	Schedule	Multi-Info	Total
Learning Rate	1e-6	28.7	10.1	5.3	19.9	29.1	11.1	6.1	20.4
	3e-6	6.9	31.6	1.0	8.4	7.1	32.0	1.0	8.6
	1e-5	7.2	27.3	0.7	8.0	6.8	30.2	1.0	8.2
	3e-5	5.9	22.7	0.3	6.6	6.0	24.2	0.4	6.8
	1e-4	8.7	3.0	3.0	6.2	8.4	3.2	3.2	6.0
	3e-4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Rank	8	30.6	2.0	3.8	19.8	31.1	2.1	4.7	20.3
	16	30.7	2.1	4.2	20	31.4	1.8	4.7	20.5
	32	31.8	2.2	4.4	20.7	33.1	2.5	5.4	21.7
	64	33.4	3.3	5.2	21.9	34.4	3.3	5.8	22.7
	128	34.8	4.2	6.4	23.1	35.1	5.1	6.3	23.4
	256	28.7	10.1	5.3	19.9	29.1	11.1	6.1	20.4

Table A10: Ablation on hyperparameters in training Llama-3.1 model with LoRA-based Direct Preference Optimization.

Hyperparam.	Value	Test Set 1 (Seen User/Unseen QA) Persona	Schedule	Multi-Info	Total	Test Set 2 (Unseen User/Unseen QA) Persona	Schedule	Multi-Info	Total
Learning Rate	1e-6	41.2	96.1	23.4	40.7	40.9	96.0	21.2	40.2
	3e-6	48.1	98.1	18.4	44.6	47	98.1	22.4	44.4
	1e-5	44.8	99.8	6.6	41.3	44.2	99.8	6.7	41.0
	3e-5	49.1	98.7	8.1	44	47.6	98.8	7.8	43.1
	1e-4	4.9	23.4	5.2	6.6	5.3	23.5	4.3	6.8
	3e-4	2.5	0.0	1.1	1.7	2.4	0.0	1.4	1.7
Rank	8.0	37.2	91.6	15.4	36.6	37	91.4	16.0	36.6
	16	39.9	93.8	18.3	38.9	38.9	93.2	17.1	38.1
	32	42.6	95.2	21.7	41.2	41.2	95.8	20.6	40.3
	64	45.0	96.2	26.0	43.4	43.7	96.2	25.7	42.5
	128	47.5	97.5	24.5	44.9	46.2	97.3	26.7	44.4
	256	48.1	98.1	18.4	44.6	47.0	98.1	22.4	44.4

Table A11: Ablation on hyperparameters in training with LoRA-based Direct Preference Optimization done after Supervised Fine-Tuning (SFT+DPO).

Meta Prompt

````
Given a task description or existing prompt, produce a detailed system prompt to guide a language model in
completing the task effectively.

# Guidelines

- Understand the Task: Grasp the main objective, goals, requirements, constraints, and expected output.
- Minimal Changes: If an existing prompt is provided, improve it only if it's simple.
  For complex prompts, enhance clarity and add missing elements without altering the original structure.
- Reasoning Before Conclusions: Encourage reasoning steps before any conclusions are reached.
  ATTENTION! If the user provides examples where the reasoning happens afterward, REVERSE the order!
  NEVER START EXAMPLES WITH CONCLUSIONS!
    - Reasoning Order: Call out reasoning portions of the prompt and conclusion parts (specific fields by name).
                       For each, determine the ORDER in which this is done, and whether it needs to be reversed.
    - Conclusion, classifications, or results should ALWAYS appear last.
- Examples: Include high-quality examples if helpful, using placeholders [in brackets] for complex elements.
   - What kinds of examples may need to be included, how many, and whether they are complex enough to benefit
     from placeholders.
- Clarity and Conciseness: Use clear, specific language. Avoid unnecessary instructions or bland statements.
- Formatting: Use markdown features for readability. DO NOT USE ``` CODE BLOCKS UNLESS SPECIFICALLY REQUESTED.
- Preserve User Content: If the input task or prompt includes extensive guidelines or examples, preserve them
                         entirely, or as closely as possible. If they are vague, consider breaking down into
                         sub-steps. Keep any details, guidelines, examples, variables, or placeholders provided
                         by the user.
- Constants: DO include constants in the prompt, as they are not susceptible to prompt injection.
             Such as guides, rubrics, and examples.
- Output Format: Explicitly the most appropriate output format, in detail. This should include length and syntax
                 (e.g. short sentence, paragraph, JSON, etc.)
    - For tasks outputting well-defined or structured data (classification, JSON, etc.)
      bias toward outputting a JSON.
    - JSON should never be wrapped in code blocks (```) unless explicitly requested.

The final prompt you output should adhere to the following structure below.
Do not include any additional commentary, only output the completed system prompt.
SPECIFICALLY, do not include any additional messages at the start or end of the prompt. (e.g. no "---")

[Concise instruction describing the task - this should be the first line in the prompt, no section header]

[Additional details as needed.]

[Optional sections with headings or bullet points for detailed steps.]

# Steps [optional]

[optional: a detailed breakdown of the steps necessary to accomplish the task]

# Output Format

[Specifically call out how the output should be formatted, be it response length, structure
 e.g. JSON, markdown, etc]

# Examples [optional]

[Optional: 1-3 well-defined examples with placeholders if necessary. Clearly mark where examples start and end,
 and what the input and output are. User placeholders as necessary.]
[If the examples are shorter than what a realistic example is expected to be, make a reference with () explaining
 how real examples should be longer / shorter / different. AND USE PLACEHOLDERS! ]

# Notes [optional]

[optional: edge cases, details, and an area to call or repeat out specific important considerations]
````

Figure A1: Meta-prompt for prompt optimization. The meta-prompt is used to generate the optimal prompt by prompting GPT-4o along with the task description. This meta-prompt is taken directly from OpenAI documentation and was created by OpenAI based on prompt engineering best practices and real-world experience (Source: https://platform.openai.com/docs/guides/prompt-generation). Generation setting: $\tau = 0.0$ (greedy).

Prompt for Profile Metadata Generation

```
Create a realistic and appropriate set of attributes (age, gender, occupation, personality, income range)
based on the given persona.

Input Persona: {persona}

Output Format:
- Age: [number]
- Gender: [gender]
- Occupation: [occupation]
- Personality: [few descriptive words]
- Income Range: [$X - $Y]
```

Figure A2: Prompt template for generating **profile metadata** in HiCUPID based on PersonaHub. The 'Input Persona' is the description for one of the 1,500 randomly-sampled individuals from PersonaHub. Generation setting: $\tau = 0.0$ (greedy).

Prompt for Schedule Metadata Generation

```
Create a set of 10 events based on the provided schedules,
ensuring each event is realistic, meaningful, and suitable for the user profile.

- Each event should align with the given datetime and be feasible within the overall schedule.
- Events should be specific, including details such as names of friends, reasons for the event,
  or other relevant context to make them memorable.
- Ensure the events are tailored to the user's profile and preferences.

User's Profile:
- Age: {age}
- Gender: {gender}
- Personality: {personality}
- Occupation: {occupation}
- Income Range: {income_range}

User's Schedules:
{datetime}

Notes:
- Ensure events are diverse and reflect the user's interests and lifestyle.
- Avoid scheduling conflicts or overly tight transitions between events.
- Include specific and personal details to make each event meaningful.
```

Figure A3: Prompt template for generating **schedule metadata** in HiCUPID. The user's schedules are explicitly conditioned on the profile metadata from above to guarantee that they are realistic, meaningful, and suitable. For each user, we generate 10 different schedules. Generation setting: $\tau = 0.0$ (greedy).

Prompt for Persona-Profile Metadata Generation

```
Select the 5 personas that best match the given user's profile from a provided list of personas.

The selected personas must align with the user's profile,
avoiding contradictions and ensuring they are realistic and plausible in real-world scenarios.
Provide the numbers of the selected personas and briefly explain how they relate to the user's profile.
Ensure the reasoning is logical and makes sense.

User's Profile:
- Age: {age}
- Gender: {gender}
- Personality: {personality}
- Occupation: {occupation}
- Income Range: {income_range}

List of Personas:
{persona}
```

Figure A4: Prompt template for generating **persona-profile combination metadata** in HiCUPID. We select five different personas that are most aligned with the user's profile, such that the multi-info questions from these combinations are realistic and logical. Generation setting: $\tau = 0.0$ (greedy).

Prompt for Persona Dialogue Generation

```
Create 10 diverse dialogues between a user and an assistant, each consisting of 20 messages
(10 turns each from the user and assistant).

Each dialogue must adhere to the following conditions:
- The dialogues should not include any form of greeting or introductory pleasantries.
- The user initiates the conversation by asking a question.
- The first question should relate to the user's given characteristics and lead into the main topic of the dialogue.
- The assistant's responses must be at least 3 sentences long, providing helpful and detailed information.
- The user reacts to the assistant's response by sharing their thoughts and asking a follow-up question.
- Around the middle of the dialogue, the user naturally shares their given characteristics while asking a question.
- The conversation must progress in a logical and natural manner.
- Each dialogue should explore a variety of topics related to the user's given characteristics.

User's Characteristics:
- The user {relation} {entity}
```

Figure A5: Prompt template for generating **persona dialogues** in HiCUPID. We create 10 different dialogues per persona to simulate 10 users sharing a common persona but having distinct interactions with the assistant. Each persona dialogue is structured to contain 10 turns. Generation setting: $\tau = 0.0$ (greedy).

Prompt for Profile Dialogue Generation

```
Create 5 dialogues between a user and an assistant, where each dialogue consists of 2 messages
(1 turn each from the user and the assistant). Each dialogue should focus on one aspect of the user's profile:
age, gender, personality, occupation, or income range.

Each dialogue must meet the following conditions:
- The user asks the assistant a question.
- The assistant provides a detailed and informative response to the user's question.
- During the question, the user shares their profile information with the assistant.
- The question should be realistic and relevant to the profile of the user.

User's Profile:
- Age: {age}
- Gender: {gender}
- Personality: {personality}
- Occupation: {occupation}
- Income Range: {income_range}
```

Figure A6: Prompt template for generating **profile dialogues** in HiCUPID. We create five profile dialogues per user, with each dialogue corresponding to one aspect of the user's profile. Each profile dialogue is structured to contain a single turn. Generation setting: $\tau = 0.0$ (greedy).

Prompt for Schedule Dialogue Generation

```
Create a dialogue between a user and an assistant consisting of two messages,
where the user informs the assistant about their schedule and makes a request such as a reminder
or adding it to a calendar, and the assistant responds diligently.

User's Schedule:
- Datetime: {datetime}
- Event: {event}
```

Figure A7: Prompt template for generating **schedule dialogues** in HiCUPID. Because each user has 10 schedules associated with him/her, we create 10 schedule dialogues per user. Each schedule dialogue is structured to contain a single turn. Generation setting: $\tau = 0.0$ (greedy).

Prompt for Persona (Single-Info) QA Pair Generation

```
Create 10 diverse sets of (question, personalized answer, general answer) based on the given user's characteristics.

Each set must meet the following conditions:
- The user asks a question related to their given characteristics without explicitly mentioning those
  characteristics.
- The personalized answer explicitly references the user's characteristics and provides a tailored response
  considering those characteristics.
- The general answer does not consider or reference the user's characteristics and provides a generic response.
- The questions should cover a variety of topics related to the user's given characteristics.

User's Characteristics:
- The user {relation} {entity}
```

Figure A8: Prompt template for generating **persona (single-info) QA** pairs in HiCUPID. As in the persona dialogue generation process, we generate 10 different persona QA pairs for 10 different users. The prompt enforces that the question does not contain cues to the user's metadata. Generation setting: $\tau = 0.0$ (greedy).

Prompt for Schedule (Single-Info) QA Pair Generation

```
Create a set of (question, personalized answer, general answer) that meets the following conditions:

- The question must involve the user notifying about a different event at the same datetime and requesting actions
  like reminders or adding it to a calendar app.
- The personalized answer must:
  - Recognize a scheduling conflict.
  - Remind the user of their previous schedule.
  - Ask if the user wants to overwrite the existing schedule.
- The general answer must:
  - Not acknowledge any prior schedule.
  - Simply respond to the user's request.

User's Schedule:
- Datetime: {datetime}
- Event: {event}
```

Figure A9: Prompt template for generating **schedule (single-info) QA** pairs in HiCUPID. The prompt enforces that the question does not contain cues to the user's metadata. Generation setting: $\tau = 0.0$ (greedy).

Prompt for Persona-Profile (Multi-Info) QA Pair Generation

```
Create a set of (question, personalized answer, general answer) that adheres to the following conditions:

- Question: The user asks a question related to their given profile and characteristics without explicitly
  mentioning them.
- Personalized Answer: The assistant provides a response that:
  1. First considers the user's profile and explicitly references it to deliver a tailored answer.
  2. Then considers the user's characteristics and explicitly references them to further customize the response.
- General Answer: The assistant provides a response that does not take the user's profile or characteristics into
  account and remains general.
- Question is related to their given profile and characteristics.
- Personalized answer must consider both the user's given profile and characteristics.

User's Profile:
- Age: {age}
- Gender: {gender}
- Personality: {personality}
- Occupation: {occupation}
- Income Range: {income_range}

User's Characteristics:
- The user {relation} {entity}
```

Figure A10: Prompt template for generating **persona-profile (multi-info) QA** pairs in HiCUPID. The prompt enforces that the question does not contain cues to the user's metadata. Generation setting: $\tau = 0.0$ (greedy).

Zero-shot Prompt in User Role (Default)

```
{
 "role": "user",
 "content": """
 You are a personalized AI assistant for the user. Follow these conditions when answering the user's questions:

 - Refer to the previous conversation history when responding to the user's question.
 - If the conversation history contains user-specific characteristics relevant to the question,
   mention those characteristics and provide a personalized response considering them.
 - If the conversation history includes user-specific information related to the question,
   reference that information and tailor your response accordingly.
 - Keep your answers concise. You must respond within a few sentences.""".strip(),
},
{
  "role": "assistant",
  "content": "Of course! How can I assist you today?",
}
```

Zero-shot Prompt in System Prompt

```
{
 "role": "SYSTEM",
 "content": """
 You are a personalized AI assistant for the user. Follow these conditions when answering the user's questions:

 - Refer to the previous conversation history when responding to the user's question.
 - If the conversation history contains user-specific characteristics relevant to the question,
   mention those characteristics and provide a personalized response considering them.
 - If the conversation history includes user-specific information related to the question,
   reference that information and tailor your response accordingly.
 - Keep your answers concise. You must respond within a few sentences.""".strip(),
}
```

Figure A11: Instruction prompt for **zero-shot inference**. The prompt placement under the "user" role is the default setting used across our experiments. We additionally analyze how placing it as a part of the system prompt affects the personalization results. Generation setting for the GPT-4o-mini model: $\tau = 0.6$, top-$p = 1.0$. Generation setting for open-source models: $\tau = 0.6$, top-$p = 1.0$, top-$k = 50$.

Few-shot Prompt in User Role (Default)

```
{
   "role": "user",
   "content": """
   You are a personalized AI assistant for the user. Follow these conditions when answering the user's questions:

   - Refer to the previous conversation history when responding to the user's question.
   - If the conversation history contains user-specific characteristics relevant to the question,
     mention those characteristics and provide a personalized response considering them.
   - If the conversation history includes user-specific information related to the question,
     reference that information and tailor your response accordingly.
   - Keep your answers concise. You must respond within a few sentences.

   # Examples

   Example 1:
     - Question: What are some classic romantic comedies to watch?
     - Personalized answer: Since you enjoy Hugh Grant's performances, you might love 'Notting Hill' and
                            'Four Weddings and a Funeral,' both of which showcase his charm.
     - General answer: Some classic romantic comedies include 'When Harry Met Sally,' 'Pretty Woman,' and
                       '10 Things I Hate About You.'

   Example 2:
     - Question: What are some effective ways to improve sprinting speed?
     - Personalized answer: Considering your profession as a sports journalist and your admiration for Usain Bolt,
                            you might find it insightful to study his training techniques and interview experts in
                            sprinting biomechanics to gain a deeper understanding.
     - General answer: To improve sprinting speed, one can focus on strength training, proper running form,
                       and consistent practice while maintaining a balanced diet and adequate rest.

   Example 3:
     - Question: Can you add a reminder for a friend's birthday party at 8 PM on April 06, 2025?
     - Personalized answer: It seems you already have a data science webinar on emerging AI trends scheduled
                            for 8 PM on April 06, 2025. Would you like to overwrite this event with the new
                            reminder for your friend's birthday party?
     - General answer: Sure, I can add a reminder for your friend's birthday party at 8 PM on April 06, 2025.

   # Notes

   - Always prioritize providing a personalized answer when possible, based on the user's characteristics and
     information from prior conversations.
   - Ensure responses are concise and within a few sentences.""".strip(),
},
{
   "role": "assistant",
   "content": "Of course! How can I assist you today?",
},
```

Figure A12: Instruction prompt for **few (1)-shot inference**. We consider a set of {persona QA, persona-profile QA, schedule QA} to be a single in-context demonstration that encompasses all question types in HiCUPID. The prompt placement under the "user" role is the default setting used across our experiments. We additionally analyze how placing it as a part of the system prompt affects the personalization results. Generation setting for the GPT-4o-mini model: $\tau = 0.6$, top-$p = 1.0$. Generation setting for open-source models: $\tau = 0.6$, top-$p = 1.0$, top-$k = 50$.

Few-shot Prompt in Systen Prompt

```
{
   "role": "SYSTEM",
   "content": """
   You are a personalized AI assistant for the user. Follow these conditions when answering the user's questions:

   - Refer to the previous conversation history when responding to the user's question.
   - If the conversation history contains user-specific characteristics relevant to the question,
     mention those characteristics and provide a personalized response considering them.
   - If the conversation history includes user-specific information related to the question,
     reference that information and tailor your response accordingly.
   - Keep your answers concise. You must respond within a few sentences.

   # Examples

   Example 1:
     - Question: What are some classic romantic comedies to watch?
     - Personalized answer: Since you enjoy Hugh Grant's performances, you might love 'Notting Hill' and
                            'Four Weddings and a Funeral,' both of which showcase his charm.
     - General answer: Some classic romantic comedies include 'When Harry Met Sally,' 'Pretty Woman,' and
                       '10 Things I Hate About You.'

   Example 2:
     - Question: What are some effective ways to improve sprinting speed?
     - Personalized answer: Considering your profession as a sports journalist and your admiration for Usain Bolt,
                            you might find it insightful to study his training techniques and interview experts in
                            sprinting biomechanics to gain a deeper understanding.
     - General answer: To improve sprinting speed, one can focus on strength training, proper running form,
                       and consistent practice while maintaining a balanced diet and adequate rest.

   Example 3:
     - Question: Can you add a reminder for a friend's birthday party at 8 PM on April 06, 2025?
     - Personalized answer: It seems you already have a data science webinar on emerging AI trends scheduled
                            for 8 PM on April 06, 2025. Would you like to overwrite this event with the new
                            reminder for your friend's birthday party?
     - General answer: Sure, I can add a reminder for your friend's birthday party at 8 PM on April 06, 2025.

   # Notes

   - Always prioritize providing a personalized answer when possible, based on the user's characteristics and
     information from prior conversations.
   - Ensure responses are concise and within a few sentences.""".strip(),
},
{
   "role": "assistant",
   "content": "Of course! How can I assist you today?",
},
```

Figure A13: Instruction prompt for **few (1)-shot inference**. We consider a set of {a persona QA pair, a multi-info (persona+profile) QA pair, a schedule QA pair} to be a single in-context demonstration that encompasses all question types in HiCUPID. The prompt placement under the "user" role is the default setting used across our experiments. We additionally analyze how placing it as a part of the system prompt affects the personalization results. Generation setting for the GPT-4o-mini model: $\tau = 0.6$, top-$p = 1.0$. Generation setting for open-source models: $\tau = 0.6$, top-$p = 1.0$, top-$k = 50$.

Survey Template for Human A/B Evaluation

```
You will be provided a user's question to an AI assistant, a user's profile,
and two sample responses from an AI-assistant.
The goal of this survey is to evaluate which sample response addresses the question better.

Examples of a user profile:
[category: 'Sports', entity: 'Soccer', sentiment: 'POS', verb: 'likes'] means that the user likes Soccer.
[category: 'Game', entity: 'Chess', sentiment: 'NEG', verb: 'dislikes'] means that the user dislikes Chess.

Choose the better response between the two responses based on the following criteria:
Logical Validity: Is the answer logically valid as a response to the question?
Personalization: Does the answer consider the user's characteristics well?

If you believe the two sample responses are of equal or comparable quality, please select "Tie."
```

Figure A14: The survey template for human A/B evaluation.

Prompt for GPT-4o A/B Evaluation on Persona and Multi-Info QA Pairs

```
Evaluate two responses (A and B) to a given question based on the following criteria:

1. Personalization: Does the response effectively consider the user's provided personal information?
2. Logical Validity: Is the response logically sound and relevant to the question?

For each criterion, provide a brief one-line comparison of the two responses and select the better response
(A, B, or Tie).

- Ensure the comparisons are concise and directly address the criteria.
- If both answers are equally strong or weak in a category, mark it as a Tie.
- Do not use bold font.

Output Format:
1. Personalization: [Brief comparison of A and B]
2. Logical Validity: [Brief comparison of A and B]
Better Response: [A/B/Tie]

Input:
- User's Personal Information:
  {metadata}
- Question: {question}
- Answer (A): {answer_a}
- Answer (B): {answer_b}
```

Prompt for GPT-4o Evaluation on Schedule QA Pairs

```
Evaluate the given answer based on the provided question and user's schedule.

The evaluation criterion is as follows:
1. Personalization: Does the answer acknowledge and appropriately consider the user's schedule?

Provide a short one-line evaluation for the criterion and indicate whether it satisfies the criterion with YES or NO.
Do not use bold font.

Output Format:
1. Personalization: [Short one-line evaluation]
Satisfaction: [YES/NO]

Example 1:
- User's Schedule:
  - Datetime: 11 AM on March 13, 2025
  - Event: Attend a networking brunch with fellow sports journalists and industry experts.
- Question: Can you add a meeting with the editorial team to my calendar for 11 AM on March 13, 2025?
- Answer: It seems you already have a networking brunch with fellow sports journalists and industry experts scheduled
          for 11 AM on March 13, 2025. Would you like to replace this event with the editorial team meeting, or should
          we consider rescheduling one of them?

1. Personalization: The answer acknowledges the existing event and offers alternatives.
Satisfaction: YES

Example 2:
- User's Schedule:
  - Datetime: 11 AM on March 13, 2025
  - Event: Attend a networking brunch with fellow sports journalists and industry experts.
- Question: Can you add a meeting with the editorial team to my calendar for 11 AM on March 13, 2025?
- Answer: Certainly, I can add the meeting with the editorial team to your calendar for 11 AM on March 13, 2025.

1. Personalization: The answer ignores the existing event in the user's schedule.
Satisfaction: NO

Input:
- User's Schedule:
  - Datetime: 9 AM on January 13, 2025
  - Event: Conduct an in-depth analysis session for an upcoming sports documentary.
- Question: Can you add a meeting with the marketing team to my calendar for 9 AM on January 13, 2025?
- Answer: Certainly, I can add the meeting with the marketing team to your calendar for the specified time.
```

Figure A15: Instruction prompts for GPT-4o evaluation. We use two different prompt templates, one to score the responses to persona and multi-info QA pairs and the other to evaluate responses to schedule QA pairs. Generation setting for GPT-4o: $\tau = 0.0$ (greedy).