Few-shot Personalization of LLMs with Mis-aligned Responses

Jaehyung Kim[1*] **Yiming Yang**[2]
[1]Yonsei University [2]Carnegie Mellon University
jaehyungk@yonsei.ac.kr

Abstract

As the diversity of users increases, the capability of providing personalized responses by large language models (LLMs) has become increasingly important. Existing approaches have only limited successes in LLM personalization, due to the absence of personalized learning or the reliance on shared personal data. This paper proposes a new approach for a **f**ew-shot p**er**sonalization of LLMs with their **mi**s-aligned responses (FERMI). Our key idea is to learn a set of personalized prompts for each user by progressively improving the prompts using LLMs, based on user profile (*e.g.*, demographic information) and a few examples of previous opinions. During an iterative process of prompt improvement, we incorporate the contexts of mis-aligned responses by LLMs, which are especially crucial for the effective personalization of LLMs. In addition, we develop an effective inference method to further leverage the context of the test query and the personalized prompts. Our experimental results demonstrate that FERMI significantly improves performance across various benchmarks, compared to best-performing baselines.[1]

1 Introduction

The recent development of large language models (LLMs) has significantly accelerated progress in various NLP tasks, and yielded real-world applications used by millions of users, such as coding assistants and chatbots (OpenAI, 2022; Team et al., 2023; Touvron et al., 2023). As the use of LLMs by diverse users in real-world applications increases, *personalization* of LLMs, *i.e.*, steering LLMs' responses towards the unique needs or preferences of individual users becomes progressively important (Glaese et al., 2022; Solaiman and Dennison, 2021). However, recent studies show that LLMs' responses are often biased toward certain groups but not suited for other diverse groups of users, and such biases cannot be fixed by providing simple instructions (Santurkar et al., 2023).

Recent work in steering the responses of LLMs can be roughly divided into two categories. One category is prompt engineering, which heuristically incorporates the user's information into the input prompts of LLMs (Salemi et al., 2023; Hwang et al., 2023). The other category focuses on learning from other users' data (Li et al., 2023; Zhao et al., 2024). Both categories have their limitations: prompt engineering for each user would be too costly as it requires exploring the extensive search space of all possible prompts to find the personalized prompt. Also, the learning-based category relies on the unrealistic assumption that personal data can be shared without violating privacy considerations.

This paper addresses those limitations by introducing a new approach, namely **F**ew-shot **Per**sonalization of LLMs with **mi**s-aligned responses (FERMI). Our high-level idea is to use LLM to progressively improve its input prompts based on a few examples of previous user opinions and profiles (*e.g.*, demographics) in an iterative process. In addition to the current prompts' scores measured on given few-shot user opinions (Yang et al., 2024), FERMI incorporates the *mis-aligned responses* (*i.e.*, LLM's responses with those prompts, which are inconsistent with given user opinions) as additional context. The contexts of mis-aligned responses include useful learning signals to update prompts such as the types of wrong predictions with the current prompts (see the empirical evidence in Section 4). Specifically, the iterative process of FERMI consists of three steps: (1) scoring the initial or current prompts with LLM, (2) updating the memory with high-scored prompts in the form of <prompt, score, context> triplets, and (3) generating new improved prompts with LLM based on the updated mem-

[*]This work is done when Jaehyung was postdoc at CMU.
[1]The code is available at https://github.com/bbuing9/Fermi.

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 11943–11974
April 29 - May 4, 2025

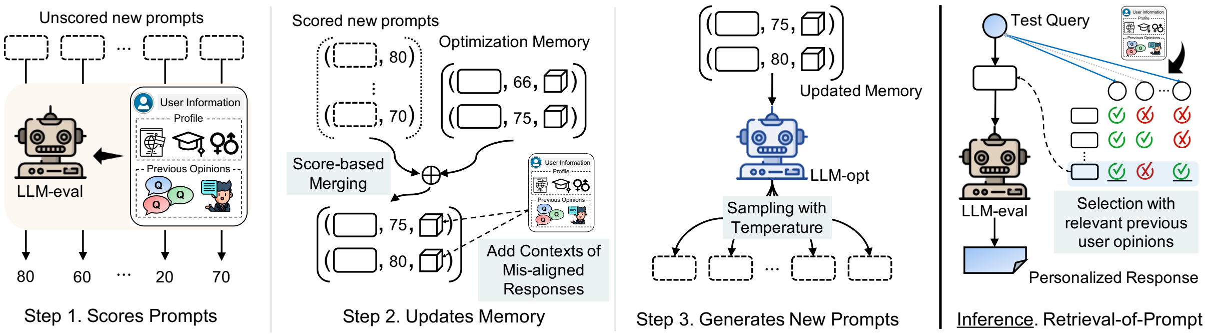


Figure 1: **An overview of FERMI.** FERMI iterates three steps to optimize the prompt from the given user information: (1) scoring new prompts, (2) updating the memory based on the score, and (3) generating new prompts (**left**). After the optimization, FERMI selectively uses the personalized prompts for the inference, via Retrieval-of-Prompt (**right**).

ory. In addition, we propose Retrieval-or-Prompt, a method to improve the inference on a given test query. Retrieval-or-Prompt selectively uses one of the personalized prompts obtained from the optimization, based on the context of the test query. An overview of FERMI is presented in Figure 1.

We demonstrate the effectiveness of FERMI for few-shot personalization of LLMs, through extensive evaluations on various tasks including question-answering (QA), classification, regression, and generation. For example, we observe that FERMI exhibited 6.8% and 4.1% average accuracy improvements on two multiple-choice QA datasets, constructed to evaluate the personalization of LLMs, compared to the previous state-of-the-art heuristic and optimization approaches, respectively. We also found that the personalized prompts produced with one LLM are also effective on other LLMs, including both API-based and open-sourced ones, which is crucial for efficient deployment in practice. In addition, our in-depth analyses reveal why FERMI is more effective than other prompting methods and what are the important features of prompts for effective personalization of LLMs. We hope our work provides useful insights for the research on LLM personalization, which becomes increasingly emerging and important for the future success of LLMs in real-world applications.

2 Related Works

Few-shot personalization of LLMs. Few-shot personalization of LLM is to align LLM's responses to a specific user with a limited number of user information such as user profile (*e.g.*, demographic information) or opinions (*e.g.*, previous responses to questions by user). To this end, one line of prior works has explored how to input given user information into LLM in a heuristical manner, *i.e.*, prompt engineering; for example, Santurkar et al. (2023) designs three different templates of input prompt. Salemi et al. (2023) leverages the retrieval system (Izacard et al., 2022) to use the given user opinions selectively. Hwang et al. (2023) shows that using both user profile and opinions is more effective. While PEFT-based personalization method has been explored (Tan et al., 2024), it's hard to be applied to recent black-box LLMs. On the other hand, another line of prior works has proposed learning from other user's data; Li et al. (2023) selects the relevant users using collaborative filtering, then learns the soft-prompt (Li and Liang, 2021) from the augmented training data from these users' data. Zhao et al. (2024) proposes to train an independent transformer module via meta-learning on several users' data. However, both approaches have their limitations; prompt engineering incurs the cost of designing the prompt, and could be limited to fully utilizing the user information due to the absence of learning. The learning-based one necessitates other users' data which inevitably incurs privacy issues. Therefore, we propose to only learn from target user's information and find the optimized (*i.e.*, personalized) prompt for that user.

Prompt optimization with LLM. As the prior works for prompt-tuning, relying on the gradient-based update (Deng et al., 2022; Lester et al., 2021; Shin et al., 2020), become inapplicable to the recent API-based LLM due to their black-box nature, other approaches have been recently explored for gradient-free prompt optimization, such as a progressive improvement using heuristic rules or LLMs (Prasad et al., 2023; Yang et al., 2024; Zhou et al., 2023). For example, Pryzant et al. (2023) receives text feedback on how to update the prompts by instructing LLM. Also, after generating initial prompts with LLMs, Zhou et al. (2023) generates a semantically similar variant of the prompts with the highest accuracies. Yang et al. (2024) iterates evaluation and generation of prompts with two LLMs, to solve the black-box optimization such as prompt optimization; Yang et al. (2024) incorporates the past generated prompts with their scores to enable the LLM for the optimization to construct new improved prompts. However, only providing the scores on training examples is insufficient to optimize the prompt for few-shot personalization of LLMs, as the context with mis-aligned responses such as the types or patterns within recursively wrong predictions can't be captured in scores. Therefore, we propose an efficient way to incorporate such context, along with an additional method to improve the inference by considering the context of the given query.

3 FERMI: Few-shot Personalization via Learning from Mis-aligned Responses

In this section, we present our framework proposed for **F**ew-shot **P**ersonalization of LLMs from **mis**-aligned responses (FERMI). We first present our problem setup in Section 3.1. Then, in Section 3.2, we present our core component that optimizes the input prompt with a given user information, by using LLM as a black-box optimizer along with the additional contexts from mis-aligned responses. Lastly, we introduce an efficient inference scheme after optimizing prompts with FERMI, by utilizing the context of a test query (Section 3.3).

3.1 Problem description

We first describe the problem setup of our interest under a question-answering (QA) scenario. Our goal is to steer LLM for a specific user using that user's information, and hence make LLM adaptively answer a given question depending on the user. Formally, let q denote the given test question and $\mathcal{M}$ denote the LLM, respectively. Next, for user u, we assume two types of user information: $U_{\mathtt{pro}}$ and $U_{\mathtt{opi}}$. $U_{\mathtt{pro}}$ indicates explicit profile of u such as demographics information (*e.g.*, region, sex, and age) or ideology (*e.g.*, political affiliation). $U_{\mathtt{opi}}$ indicates N few-shot previous opinions by u, which has the form of QA pairs, *i.e.*, $U_{\mathtt{opi}} = \{(q_i, a_i)\}_{i=1}^{N}$ where q_i is a previously asked question and a_i is an opinion (answer) by the user. Then, for given test question q, our goal is to predict the answer a, which would be generated by user u, through LLM $\mathcal{M}$ using both $U_{\mathtt{pro}}$ and $U_{\mathtt{opi}}$. The heuristic design of input prompt p to incorporate such user information has been previously explored (Hwang et al., 2023; Santurkar et al., 2023), *i.e.*, prediction $\widehat{a}$ is obtained by conditioning $\mathcal{M}$ with p constructed using $U_{\mathtt{pro}}$ and $U_{\mathtt{opi}}$:

$$\widehat{a}(\mathrm{p}) = \mathcal{M}(q; \mathrm{p}). \quad (1)$$

However, heuristically designed prompts are limited to fully exploit the given user information; for example, the personalization of LLM was better with few user opinions (*e.g.*, 3 or 8) compared to using all opinions in $u_{\mathtt{opi}}$ (Hwang et al., 2023). Therefore, we tackle this limitation by finding personalized prompts that align LLM to the user, through direct learning from given user information.

3.2 Effective prompt optimization with LLM from contexts of mis-aligned responses

To mitigate the difficulties from the large scale and black-box nature of recent LLMs, we instead optimize input prompts to learn from user information. It is motivated by the recent work (Yang et al., 2024) that uses two LLMs, $\mathcal{M}$ and $\mathcal{M}_{\mathtt{opt}}$, to solve black-box optimization, where $\mathcal{M}_{\mathtt{opt}}$ denotes another LLM used for the optimization. Specifically, our key idea is incorporating the contexts of *mis-aligned* responses (*i.e.*, QAs in $U_{\mathtt{opi}}$ that $\mathcal{M}$

I want to find the text that could make you a personalized answer for the question about **Gun**, based on the given personal information … If your response is identical to the person's, then you get a score of 1; otherwise, you get a score of 0. Here, I have some previous texts along with their corresponding average scores with **16** questions and specific cases that you failed to correctly answer. The texts are arranged in ascending order …

Optimization Memory (**varied**)

text: **Prompt #1**
score: **Score of prompt #1**
failure cases:
<1>
Question: How often, if ever, do you visit websites about guns, hunting or other shooting sports
Answer choices: A. Often B. Sometimes C. Hardly ever D. Never
Answer: D
Your response: C
<3> … <10> … <11> … <14> … <16>

text: **Prompt #2**
score: **Score of prompt #2**
failure cases: **[1, 3, 10, 11, 16] and 1 additional examples that was correctly predicted with the first text**

⋮

The following exemplars show how to apply your text: you replace <INS> in each input with your text, then read the input and give an output …

Few-shot Demonstration (fixed)

[1]
<INS>
Question: How often, if ever, do you visit websites about guns, hunting or other shooting sports
Answer choices: A. Often B. Sometimes C. Hardly ever D. Never
Answer: D
[2] … [3] … [4] …

Write your new text that is different from the old ones and has a score as high as possible. Write the text in square brackets.

Figure 2: **Prompt example.** Example of input prompt for $\mathcal{M}_{\mathtt{opt}}$ to generate new prompts, composed of fixed input prompt $\mathrm{p}_{\mathtt{opt}}$ (including fixed few-shot demonstrations) and optimization memory M^t (Eq. 5) on OpinionQA dataset. A full version is in Appendix B.3.

incorrectly predict with current prompts) during the optimization, instead of only using scores of the prompts (*e.g.*, average accuracy of the prediction by $\mathcal{M}$ on $U_{\texttt{opi}}$). As the contexts of mis-aligned responses include useful learning signals such as types or patterns of common wrong predictions, they could be effective in learning how to improve the prompts.

FERMI starts with an initial prompt set $\mathrm{P}^0 = \{\mathrm{p}^0\}$ and we adopt heuristically designed prompts in previous works (Hwang et al., 2023; Santurkar et al., 2023) for p^0. Specifically, we use the user's explicit profile $U_{\texttt{pro}}$ when it is available; thereby we fully utilize the given user information. If not, we adopt vanilla prompt that instructs QA without user information (see details in Appendix B.3).

Then, FERMI iterate the following 3 steps: *(1) Score Prompts*, *(2) Update Memory*, and *(3) Generate New Prompts*. See Figure 1 for the illustration.

∘ *Step 1: Score Prompts*. We first calculate the score s_k of each prompt $\mathrm{p}_k \in \mathrm{P}^t$, by obtaining the predictions from $\mathcal{M}$ under p_k and evaluating them using the user's previous answers:

$$s_k = \sum\nolimits_{(q_i,a_i)\sim U_{\texttt{opi}}} \mathrm{s}\big(a_i, \widehat{a}_i(\mathrm{p}_k)\big)/N, \quad (2)$$

where $\widehat{a}_i(\mathrm{p}_k) = \mathcal{M}(q_i; \mathrm{p}_k)$. Here, $\mathrm{s}(\cdot,\cdot)$ is a specific metric to evaluate the prediction (*e.g.*, accuracy). During this calculation of the score s_k of the prompt p_k, we also collect pair of QA pairs $U^k_{\texttt{opi}}$ that $\mathcal{M}$ yields *mis-aligned response* $\widehat{a}_i$ to the user's answer a_i, under p_k:

$$U^k_{\texttt{opi}} = \{(q_i,a_i) \in U_{\texttt{opi}} | \mathrm{s}\big(a_i, \widehat{a}_i(\mathrm{p}_k)\big) < \tau\}, \quad (3)$$

where τ is a threshold to judge the mis-alignment; for example, we set $\tau = 0.5$ when we use the correctness of prediction as the score $\mathrm{s}(\cdot,\cdot)$.

∘ *Step 2: Update Memory*. Next, we construct an optimization memory M^t, which is used for the input of $\mathcal{M}_{\texttt{opt}}$ to generate new improved prompts, by providing the information of well-performing prompts through the contexts of their mis-aligned responses. To be specific, the optimization memory $M^t = \{(\mathrm{p}_l, s_l, c_l)\}_{l=1}^{L}$ is constructed by selecting top-L prompts among P^t and M^{t-1} (where $M^0 = \emptyset$), according to their scores (Eq. 2). Here, we present the triplets in M^t in ascending order, *i.e.*, $s_l < s_{l'}$ when $l < l'$, and provide the varied context c_l depending on l. Specifically, for $l = 1$, we construct c_l by concatenating QAs and mis-aligned responses by $\mathcal{M}$ under p_l on $U^l_{\texttt{opi}}$:

$$c_l = \texttt{Concat}\{\big(i, q_i, a_i, \widehat{a}_i(\mathrm{p}_l)\big) | (q_i,a_i) \in U^l_{\texttt{opi}}\}. \quad (4)$$

In Figure 2, the texts corresponding to c_1 are highlighted in blue. For other cases (*i.e.*, $l \neq 1$), instead of the enumeration like c_1, we construct the context c_l with (i) the indices of *common* mis-aligned QA pairs between p_l and p_1, and (ii) the number of *newly* mis-aligned QAs by p_l compared to p_1 (see green texts in Figure 2 for an example). Through the presented indices in c_l, $\mathcal{M}_{\texttt{opt}}$ can directly access the mis-aligned QA pairs by referring c_1, and one can avoid unnecessary complexity of c_l and cost from the long input to $\mathcal{M}_{\texttt{opt}}$. Additionally, the number of newly mis-aligned ones offers further insight into whether p_l has improved, which can't be captured by the common mis-aligned ones.

∘ *Step 3: Generate New Prompts*. With the updated memory M^t, we generate K new improved prompts $\mathrm{P}^{t+1} = \{\mathrm{p}^{\texttt{new}}_k\}_{k=1}^{K}$ by prompting $\mathcal{M}_{\texttt{opt}}$ to generate the new and high-scored prompts:

$$\mathrm{p}^{\texttt{new}}_k = \mathcal{M}_{\texttt{opt}}(M^t; \mathrm{p}_{\texttt{opt}}), \quad (5)$$

where $\mathrm{p}_{\texttt{opt}}$ is a fixed input prompt for $\mathcal{M}_{\texttt{opt}}$ to generate new prompts, and we use a random sampling with temperature to generate diverse new prompts from $\mathcal{M}_{\texttt{opt}}$. Figure 2 presents the example of the overall input of $\mathcal{M}_{\texttt{opt}}$ to generate new prompts, which is constructed with M^t and $\mathrm{p}_{\texttt{opt}}$.

Then, we go back to Step 1 with P^{t+1} and iterate these 3 steps for T times. After that, we obtain the optimized (*i.e.*, personalized) prompts $\mathrm{P}^T = \{\mathrm{p}^T_k\}_{k=1}^{K}$ for the user u.

3.3 Effective inference by Retrieval-of-Prompt

After T iterations of the optimization procedure, FERMI outputs K unique personalized prompts $\mathrm{P}^T = \{\mathrm{p}^T_k\}_{k=1}^{K}$. Therefore, for a given test question q, one needs to determine which prompt to apply. Selecting the prompt with the highest score, *i.e.*, $k^* = \arg\max_k s_k$ (Eq. 2), would be a straightforward way. However, our intuition is that better selection is possible if we utilize the context of the test question q as additional information. To this end, we propose to select the input prompt with the highest score on the subset of $U_{\texttt{opi}}$, which only consists of the previous questions highly relevant to q. Formally, we first measure the relevance r between q and previous question q_i:

$$R(q, U_{\texttt{opi}}) = \{r(q, q_i) | q_i \in U_{\texttt{opi}}\}. \quad (6)$$

For the relevance r, we use the cosine similarity between the embeddings of questions, extracted by the sentence encoder (Reimers and Gurevych, 2019). Then, we select top-$\tilde{N}$ questions according to the calculated relevance and construct the subset $U^q_{\texttt{opi}}$ with those questions. Lastly, we choose the input prompt $\mathrm{p}^* = \mathrm{p}^T_{k^*}$ based on the score on $U^q_{\texttt{opi}}$, which were already calculated, and use the prediction $\widehat{a}(\mathrm{p}^*)$ by $\mathcal{M}$:

$$k^* = \arg\max_k s^T_k(U^q_{\texttt{opi}}), \tag{7}$$

where $s^T_k(U^q_{\texttt{opi}}) = \sum_{(q_i,a_i)\sim U^q_{\texttt{opi}}} s\big(a_i, \widehat{a}_i(\mathrm{p}^T_k)\big)/\tilde{N}$. Figure 1 illustrates the overview of FERMI. We note that a full version of the prompts and examples of personalized prompts are presented in Appendixes B and D, respectively.

4 Experiments

In this section, we design our experiments to investigate the following questions:

- How does FERMI perform compare to other personalization methods? (Tables 1 and 2)
- Is the optimized prompt with FERMI from one LLM transferable to different LLMs? (Table 3)
- What is the effect of each component in FERMI? (Table 4)
- Why optimized prompt by FERMI is more effective than other prompts? (Table 5)

4.1 Setups

First, we describe our experimental setups. More details are presented in Appendix B.

Datasets. For the experiments, we first use two multiple-choice QA datasets proposed to measure the steerability of LLMs for specific users (or social groups): *OpinionQA* (Santurkar et al., 2023) and *GlobalOpinionQA* (Durmus et al., 2023). For OpinionQA, we use a subsampled split released by Hwang et al. (2023), which consists of 10.5k and 15.8k training and test QA pairs across 525 users and 15 topics, respectively. For GlobalOpinionQA, since the dataset originally included the answer distribution by multiple respondents in the same country, we converted it to have a single answer by selecting the choice with the highest probability. It results in 920 training and 1,317 test QA pairs across 46 countries. We consider each country as a specific user. Next, we use three additional tasks, *LaMP*$_{\texttt{tag}}$ (classification), *LaMP*$_{\texttt{rate}}$ (regression), *LaMP*$_{\texttt{title}}$ (generation), from a recent benchmark proposed for personalization of LLMs (Salemi et al., 2023). LaMP$_{\texttt{tag}}$ is a 15-way classification task where an input is a movie description and a label is a movie tag, and LaMP$_{\texttt{rate}}$ is a regression task where an input is a user review and a label is an integer rating (1-5). LaMP$_{\texttt{title}}$ is a generation task where the input is the abstract of the paper and the goal is generating the personalized title of the paper based on the given abstract. We construct these datasets by subsampling from their original validation split, which results in 1,000 training and 1,500 test QA pairs across 50 users for each dataset. On average across four datasets, for each user, 20 training QAs as previous opinions and specific profile are given, and then 30 test QAs are used to evaluate. Following Salemi et al. (2023), we report average test accuracy (Acc), mean absolute error (MAE), and Rouge-L score for *LaMP*$_{\texttt{tag}}$, *LaMP*$_{\texttt{rate}}$, and *LaMP*$_{\texttt{title}}$, respectively.

Baselines. We compare FERMI against extensive baselines as follows: (1) *Uniform*: expected performance when the prediction is made uniformly at random. (2) *Vanilla*: answers the question with LLMs without any user information. (3) *Profile*: constructing prompt using all available user profiles (Santurkar et al., 2023; Hwang et al., 2023) such as demographics or nationality. (4) *Few-shot*: retrieving relevant previous questions and opinions, then append them to the prompt (Hwang et al., 2023; Salemi et al., 2023). Following (Salemi et al., 2023), we consider BM25 (Robertson et al., 2009) and Contriever (Izacard et al., 2022) for the retriever models. The number of retrieved profiles is determined among {3, 8, all} with validation performance. (5) *All Info*: using both explicit profiles and retrieved previous QAs to construct prompt (Hwang et al., 2023). We use the retrieval with the best performance in *Few-shot*.[2] (6) Optimization by PROmpting (*OPRO*; Yang et al. (2024)): optimizing input prompt using both user profiles and previous opinions using LLMs. Here, all of the previous opinions are utilized during the optimization. In the experiments, the prompt with the best training score is selected for the test.

Implementation details. We use three recent state-of-the-art LLMs for the prediction LLM $\mathcal{M}$ for the experiments: ChatGPT (`gpt-3.5-turbo-0613`) (OpenAI, 2022), GPT-4 (`gpt-4-turbo-1106`) (OpenAI, 2023), Mistral-7B-Instruct-v0.2 (Jiang et al., 2023) and LLaMA2-chat-70B (Touvron et al., 2023). For

[2] In the case of OpinionQA, we additionally consider the retrieved indices originally included by (Hwang et al., 2023).

Table 1: **Main result on multiple-choice QA datasets.** Test accuracy of ChatGPT over the different methods on OpinionQA (OpQA) and GlobalOpinionQA (GOQA). The best and second best scores are highlighted in **bold** and underline, respectively.

Methods	Datasets (Metric)	
	OpQA (Acc.)	GOQA (Acc.)
Uniform	34.2	31.4
Vanilla	45.5	62.8
Profile	48.1	66.1
Few-shot$_{\texttt{bm25}}$	49.8	59.1
Few-shot$_{\texttt{cont}}$	49.3	61.2
All Info	48.6	62.3
OPRO	<u>50.2</u>	<u>71.1</u>
FERMI	**54.6**	**74.8**

Table 2: **Main result on LaMP Benchmark.** Test performance of ChatGPT over the different methods on LaMP benchmarks. Test accuracy (Acc (↑)), mean absolute error (MAE (↓)), and Rouge-L (↑) score are used, respectively. The best and second best scores are highlighted in **bold** and underline, respectively.

Methods	Datasets (Metric)		
	LaMP$_{\texttt{tag}}$ (Acc.)	LaMP$_{\texttt{rate}}$ (MAE)	LaMP$_{\texttt{title}}$ (Rouge-L)
Uniform	6.7	1.65	-
Vanilla	36.1	0.62	0.394
Few-shot$_{\texttt{bm25}}$	35.9	0.40	<u>0.411</u>
Few-shot$_{\texttt{cont}}$	<u>36.2</u>	<u>0.36</u>	0.406
OPRO	34.3	0.57	0.406
FERMI	**37.8**	**0.34**	**0.419**

$\mathcal{M}$, we use a temperature of 0.0 when calling the API or greedy decoding for LLaMA, to remove the effect of random sampling. For the optimization LLM $\mathcal{M}_{\texttt{opt}}$, we always use GPT-4, as the prompt optimization based on the memory (Eq. 5) requires complex reasoning capability (See Appendix A), with a temperature of 1.0. For OPRO and FERMI, we use fixed values of $K = 4$, $L = 5$, and $T = 10$. Also, with previous user opinions in $U_{\texttt{opi}}$, 80% is used for optimization and 20% is used as few-shot demonstrations in $\mathrm{p}_{\texttt{opt}}$. We set τ, a threshold to define the mis-aligned responses (Eq. 2), 0.2 only for *LaMP*$_{\texttt{title}}$ and 1.0 for other tasks. To obtain sentence embeddings for Retrieval-of-Prompt, we use the sentence encoder with MPNet (Song et al., 2020) showing the best performance.[3] We use a fixed $\tilde{N} = 3$ for Retrieval-of-Prompt.

[3]Following the results in `https://www.sbert.net`

Table 3: **Transferability of FERMI.** Test accuracy of two LLMs on GlobalOpinionQA. For Few-shot, we use Contriever which shows higher accuracy in Table 1. For OPRO* and FERMI*, prompts optimized on ChatGPT are directly used. The best and second best scores are highlighted in **bold** and underline, respectively.

Methods	Mistral-v0.2	LLaMA-2	GPT-4
Vanilla	57.6	62.4	56.7
Profile	67.4	<u>65.5</u>	77.7
Few-shot	70.1	60.5	68.9
All Info	<u>71.7</u>	65.1	<u>78.2</u>
OPRO*	66.4	64.5	76.7
FERMI*	**72.4**	**68.9**	**84.8**

4.2 Main results

Table 1 summarizes the experimental results on two different multiple-choice QA datasets, under ChatGPT. First, it is observed that augmenting the user information into the input prompt is effective in improving the accuracies of LLMs, but the effectiveness could be varied. For example, retrieving relevant user opinions is more effective than using the user profile for OpinionQA (49.8% vs. 48.1%), but it's vice versa in GlobalOpinionQA (61.2% vs. 66.1%). It is due to the difference between datasets, as each user is asked multiple questions on the same topic in OpinionQA while GlobalOpinionQA asks the broader topics; this result also reveals the necessity of the learning-based prompt optimization approach. From the results of OPRO and FERMI, one can observe that the optimization-based approach is actually effective, and the proposed method significantly improves it. To be specific, FERMI exhibits 6.75% average accuracy improvement compared to the previous prompting method. Furthermore, compared to the existing optimization method, FERMI exhibits 4.05% accuracy improvement in the average. In Figure 3, we additionally present detailed results on OpinionQA, a topic-wise accuracy from four representative baselines selected based on average accuracy. Here, FERMI consistently shows better performance than other baselines across all topics, which further demonstrates the effectiveness of FERMI for the personalization of LLMs.

Next, Table 2 summarizes the experimental results on LaMP$_{\texttt{tag}}$, LaMP$_{\texttt{rate}}$, and LaMP$_{\texttt{title}}$. We note that these datasets do not include explicit user profiles; hence, we exclude both *Profile* and *All Info* for the baselines. Here, it is noteworthy that the effectiveness of OPRO is significantly degraded, as the given task becomes more challenging to solve

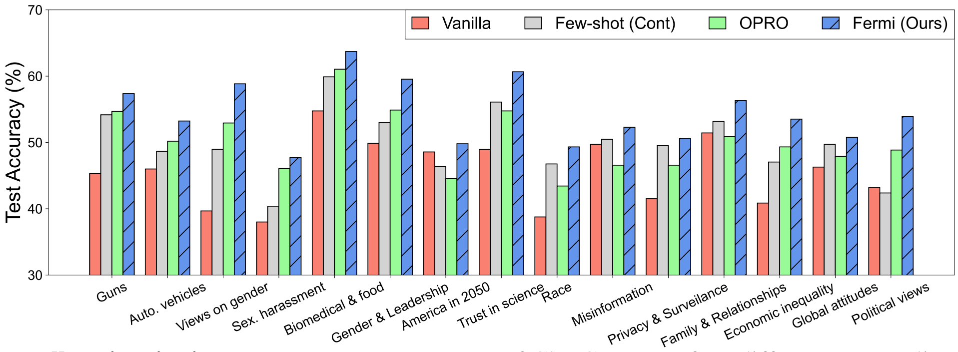


Figure 3: **Overall topic-wise improvement.** Test accuracy of ChatGPT over four different personalization methods on OpinionQA. Detailed results are presented in Appendix C.

(*e.g.*, the average number of answer choices: 3.96 for GlobalOpinionQA vs. 15 for LaMP$_{\texttt{tag}}$). Nevertheless, FERMI is consistently effective and outperforms the other baselines; for example, FERMI exhibits 17.7% and 4.0% relative improvement in average, compared to the vanilla and the previous best baseline, respectively.

4.3 Analyses with FERMI

In this section, we provide additional analyses of FERMI with the experiments on GlobalOpinionQA. More analyses are presented in Appendix A.

Transferability of the optimized prompt. Here, we provide additional experiments to verify the transferability of the learned prompt with our method. To be specific, we first save the optimized prompts under ChatGPT as LLM for evaluation (Eq. 1), which are used in Table 1. Then, we directly apply these prompts to two different types of LLMs (Mistral-7B-Instruct-v0.2, LLaMA-2-chat-70B, and GPT-4), without additional optimization as same as applying heuristically designed prompts. From Table 3, one can observe that the transferred prompts from FERMI significantly outperform the baseline prompting methods on both LLMs; for example, it exhibits 3.6% accuracy improvement compared to the best-performing baseline on average. We remark that the prompts from OPRO are even less effective than the existing baseline, which further shows the advantages of FERMI in learning the well-generalized personalized prompt.

Ablation study. To validate the effectiveness of the proposed component of FERMI in Section 3, we perform the ablation experiments by decomposing our framework into three different components: (1) including QAs that have mis-alinged responses with the initial presentation and referring via common indices (*Add*$_{\texttt{Mis}}$), (2) noting the number of QAs with new mis-aligned responses (*Add*$_{\texttt{Num}}$), and (3) Retrieval-of-Prompt for a test query (*RoP*). As shown in Table 4, all components progressively improve the few-shot personalization of LLMs. Especially, it is observable that efficiently providing the context of mis-aligned QAs during the optimization is mostly crucial for the improvement. Next, providing the number of new mis-aligned QAs makes additional improvement, as it can provide information about the effectiveness of the given prompt, which is not captured by commonly mis-aligned QAs. Lastly, for a test query, retrieving the most relevant prompt is more effective than selecting

Table 4: **Ablation study of FERMI.** Test accuracy of ChatGPT on GlobalOpinionQA with different configurations of the proposed components in FERMI.

Methods	Add$_{\texttt{Mis}}$	Add$_{\texttt{Num}}$	RoP	Acc
OPRO	✗	✗	✗	71.1
	✓	✗	✗	73.7
	✓	✓	✗	74.2
FERMI	✓	✓	✓	74.8

Table 5: **In-depth analyses about prompts for personalization.** Training and test accuracies of ChatGPT over the different methods on GlobalOpinionQA. Training accuracy is measured by given user opinions $u_{\texttt{opi}}$.

Methods	Acc$_{\texttt{train}}$	Acc$_{\texttt{test}}$
Vanilla	62.5	62.8
Profile	67.9	66.1
Few-shot$_{\texttt{top3}}$	-	61.2
Few-shot$_{\texttt{all}}$	95.2	56.3
Few-shot$_{\texttt{bott3}}$	-	45.8
Few-shot$_{\texttt{format}}$	70.2	66.4
FERMI$_{\texttt{irrel}}$	80.2	73.8
FERMI	81.4	74.8

<INS> Your answer should be a single alphabet among given answer choices:

Question: Has each of the following gotten better or worse over the last five years in our country: the availability of food in stores.

Answer choices:
A. Better B. Worse C. Has not changed

Answer:

[1]. Question … Answer choices … Answer …
[2]. Question … Answer choices … Answer …
⋮
[N]. Question … Answer choices … Answer …

Based on the above previous questions and answers, choose the proper answer to the given question among the given answer choices as if you currently reside in Bulgaria.

To adeptly navigate multiple-choice questions in the context of Bulgarian popular opinion, you must adeptly tune into the heartstrings of the national zeitgeist. Consider the collective Bulgarian experience, informed by historical nuances, current social dynamics, and the pulse of political life that shapes the everyday reality in Bulgaria. Set personal biases aside in favor of an answer that resonates with the collective beliefs and societal norms that most Bulgarians would agree with in the context of the question. Your chosen response should epitomize the conventional wisdom or the dominant perceptions that typify the Bulgarian sentiment relevant to the topic.

Figure 4: **Qualitative comparison.** Example prompts from All Info (**middle**) and FERMI (**bottom**) for the specific question (**top**) from GlobalOpinionQA. Prompt is inserted to <INS>. More examples are in Appendix D.

with the highest training score, as it successfully utilizes the context of the test query.

Features of good input prompts for personalization. In Table 5, we further conduct the experiments to answer the following question: *what features make good personalized prompts for LLMs?* First, we claim that the relevance of the prompt to the test query is crucial; for example, Few-shot$_{\texttt{top3}}$, Few-shot$_{\texttt{all}}$, and Few-shot$_{\texttt{bott3}}$ are different prompting methods by retrieving the 3 mostly relevant, all 20, and 3 mostly irrelevant previous opinions, respectively. Here, it is observable that test accuracy largely degrades when a portion of irrelevant opinions increases. Similarly, when we retrieved the most irrelevant prompt (FERMI$_{\texttt{irrel}}$), *i.e.,* take $\arg\min$ in Eq. 7, accuracy of FERMI is also decreased.

Second, providing the user information with the proper format for LLMs is important. As shown in Figure 4, the optimized prompt by FERMI is a detailed instruction consisting of multiple sentences that condense the lessons from the user opinions and LLM's mis-aligned responses. In contrast, the previous prompt used to incorporate previous opinions is based on the specific form, which is harder to follow by LLMs. To verify the importance of the format, we convert the enumeration of all QAs (by Few-shot$_{\texttt{all}}$) into the instruction of multiple sentences (denoted by Few-shot$_{\texttt{format}}$), by prompting GPT-4 using the optimized prompts by FERMI as reference. Interestingly, this format conversion shows significant improvement (56.3% $\rightarrow$ 66.4%) while it is still underperforming FERMI.

Lastly, effectively distilling the given user information is important. As shown in Table 5, the prompting method with higher accuracy on previous user opinions $U_{\texttt{opi}}$ (*i.e.*, training accuracy) has a higher test accuracy for that user as well, except Few-shot$_{\texttt{all}}$ which can directly access $U_{\texttt{opi}}$. In this aspect, FERMI shows a clear advantage compared to the previous prompting optimization method; as shown in Figure 5, FERMI more effectively optimizes the prompt and achieves higher training accuracy than OPRO. These results indicate that finding a proper way to condense and incorporate the user information to design input prompts is crucial, and FERMI achieves this by using the context of mis-aligned responses.

Overall, designing personalized prompts satisfying these three properties (relevancy to test query, proper format, and effective distillation of user information) is challenging, but FERMI effectively accomplishes this goal.

5 Conclusion

In this work, we propose FERMI, a simple yet effective framework for improving the few-shot personalization of LLMs. Our key idea is to optimize the input prompt by learning from the user information; we propose an efficient way to incorporate contexts of mis-aligned responses by LLMs during the optimization, and a retrieval approach to select the optimized prompt relevant to test query. The effectiveness of FERMI is demonstrated by results on various personalization tasks and LLMs.

Limitations

One possible limitation of FERMI is its computational cost. As discussed in Appendix A, our framework necessitates a strong LLM as the optimization of prompt requires complex reasoning capability. If we substitute $M_{\texttt{opt}}$ from GPT-4 to ChatGPT, it can't properly optimize the input prompt; see Figure 5 and we remark the similar observation was in the previous work (Yang et al., 2024).

Nevertheless, we would like to emphasize that the proposed FERMI is not just a simple consequence of more computations and costs. Compared to OPRO (Yang et al., 2024), another computationally intensive method for accuracy improvement, FERMI significantly outperforms it, *i.e.*, FERMI is an even more efficient way to increase performance. For instance, as shown in Figure 6(c), FERMI achieves much higher accuracy than OPRO, (73.0% v.s. 71.1% on GlobalOpinionQA), even with half of the previous question and user's responses and 1.4% less cost per user (0.33$ v.s. 0.34$). The effectiveness of FERMI for optimizing personalized prompts is from more effective optimization by extracting the effective learning signal from mis-aligned responses. On the other hand, one can directly control the computational cost and performance, by varying the number of iterations (T) which linearly increases the cost with improved performance. Regarding this, we remark that 4 iterations of optimization with FERMI yield similar results to the 10 iterations of optimization with OPRO (see Figure 5).

We further remark that the overall cost to use our framework will decrease while preserving its effectiveness, as the cost of using LLM with strong reasoning capability is continuously reduced. Currently, more than half of the overall cost occurs to use strong LLM (*e.g.*, GPT4) for generating the improved prompts, and this choice is inevitable as this task requires the complex reasoning capability for LLM. However, as shown in Table 6, we empirically observe that the recently released strong yet efficient LLM (GPT-4o-mini) can successfully optimize the prompt and yield a comparable performance with GPT-4, although it only requires 0.5% input token and 1.0% output per token price compared to GPT-4.

At the same time, as we demonstrated in the experiments, the personalized prompts from our method are well-transferrable to other LLMs that are not used during optimization (Table 3), could be continuously updated with enlarged data through the user interactions (Table 9), and also reusable to convert previous prompts to have the proper format for LLMs (Table 5). Therefore, we believe that our approach could be an even more efficient way for personalization compared to the heuristical design of the prompt, after the consumption of the cost at the initial optimization.

Broader impact and ethical implications

We strongly believe that FERMI can provide a strong positive impact in real-world applications that require personalized responses for the given user, *e.g.*, search engines or chatbots. We expect that our framework would be especially beneficial for the users belonging to under-populated social groups, since LLMs are known to follow the knowledge or opinion of the major population within pre-trained data (Kandpal et al., 2023; Santurkar et al., 2023). In contrast, there also exists some potential negative impacts. Since our framework needs to provide personal information to LLMs (mostly through API), it has a potential privacy risk when the provider of LLMs does not follow the safeguard and collects the given information. In addition, as our framework didn't filter out the resulting prompts separately, it can include the prompts that have socially negative impacts, *e.g.,* jailbreak of LLMs (Chao et al., 2023). We believe that the incorporation of an additional filtering step could be a solution to this problem (Xie et al., 2023).

References

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. 2023. Jailbreaking black box large language models in twenty queries. In *R0-FoMo: Robustness of Few-shot and Zero-shot Learning in Large Foundation Models*.

Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric Xing, and Zhiting Hu. 2022. Rlprompt: Optimizing discrete text prompts with reinforcement learning. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

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

Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, et al. 2022. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*.

EunJeong Hwang, Bodhisattwa Prasad Majumder, and Niket Tandon. 2023. Aligning language models to user opinions. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. In *Transactions on Machine Learning Research (TMLR)*.

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*.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

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

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

OpenAI. 2022. Introducing chatgpt. `https://openai.com/blog/chatgpt`.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

PewResearch. Writing survey questions. `https://www.pewresearch.org/our-methods/u-s-surveys/writing-survey-questions`.

Archiki Prasad, Peter Hase, Xiang Zhou, and Mohit Bansal. 2023. Grips: Gradient-free, edit-based instruction search for prompting large language models. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Reid Pryzant, Dan Iter, Jerry Li, Yin Tat Lee, Chenguang Zhu, and Michael Zeng. 2023. Automatic prompt optimization with" gradient descent" and beam search. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Anastasia Razdaibiedina, Yuning Mao, Rui Hou, Madian Khabsa, Mike Lewis, and Amjad Almahairi. 2023. Progressive prompts: Continual learning for language models. In *International Conference on Learning Representations (ICLR)*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

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

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the International Conference on Machine Learning (ICML)*.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Irene Solaiman and Christy Dennison. 2021. Process for adapting language models to society (palms) with values-targeted datasets. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems (NeurIPS)*.

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*.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. 2022. Learning to prompt for continual learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yueqi Xie, Jingwei Yi, Jiawei Shao, Justin Curl, Lingjuan Lyu, Qifeng Chen, Xing Xie, and Fangzhao Wu. 2023. Defending chatgpt against jailbreak attack via self-reminders. *Nature Machine Intelligence*, pages 1–11.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. 2024. Large language models as optimizers. In *International Conference on Learning Representations (ICLR)*.

Siyan Zhao, John Dang, and Aditya Grover. 2024. Group preference optimization: Few-shot alignment of large language models. In *International Conference on Learning Representations (ICLR)*.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023. Large language models are human-level prompt engineers. In *International Conference on Learning Representations (ICLR)*.

Qi Zhu, Bing Li, Fei Mi, Xiaoyan Zhu, and Minlie Huang. 2022. Continual prompt tuning for dialog state tracking. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Table 6: **Different LLM for $\mathcal{M}_{\texttt{opt}}$.** Test accuracy of ChatGPT over the baseline methods and FERMI with the different LLMs (GPT-4o-mini and GPT-4) for generating prompts on GlobalOpinionQA.

Models	Methods			
	Vanilla	Profile	$\text{FERMI}_{\texttt{mini}}$	FERMI
ChatGPT	62.8	66.1	74.2	74.8

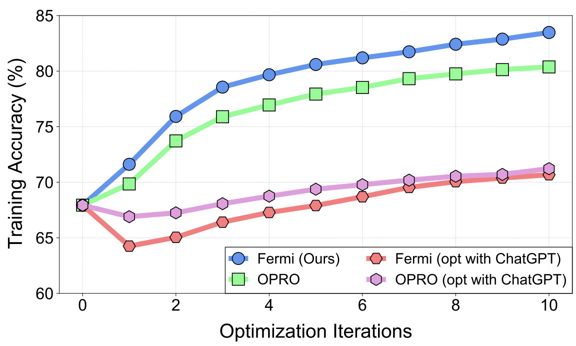


Figure 5: **Optimization trajectory under different LLMs for $\mathcal{M}_{\texttt{opt}}$.** Average training accuracies on GlobalOpinionQA across optimization iterations ($T = 10$) under OPRO and FERMI.

A More Analyses with FERMI

In this section, we provide more analyses of FERMI in addition to the analyses in Section 4.3.

GPT-4 for optimization $\mathcal{M}_{\texttt{opt}}$. As denoted in Section 4.1, we commonly use GPT-4 for LLM $\mathcal{M}_{\texttt{opt}}$ to generate new prompts from the optimization memory (Eq. 5) for all the experiments in Section 4. To validate this design choice, we conduct the experiments by substituting GPT-4 with ChatGPT $\mathcal{M}_{\texttt{opt}}$ in both OPRO and FERMI. Figure 5 is the optimization trajectory in terms of training accuracy (*i.e.*, average accuracy of the prediction by $\mathcal{M}$ on previous user opinions). Here, one can observe that both OPRO and FERMI suffer in optimizing the prompt when we use ChatGPT as $\mathcal{M}_{\texttt{opt}}$, similar to the previous observation (Yang et al., 2024); it reveals that generating the improved prompts from the optimization memory with previous prompts, scores, and contexts requires complex reasoning capability. Therefore, using a strong LLM with sufficient capability to optimize input prompts, such as GPT-4, is necessary. But, we remark that one can substitute GPT-4 with cheaper LLMs with a sufficient reasoning capability (*e.g.*, GPT-4o-mini), as shown in Table 6

Table 7: **FERMI only using GPT-4.** Test accuracy of GPT-4 over the different prompting methods on GlobalOpinionQA. For Few-shot, we use Contriever which shows higher accuracy in Table 1. For OPRO* and FERMI*, prompts optimized on ChatGPT are directly used. The best and second best scores are highlighted in **bold** and <u>underline</u>, respectively.

Methods	GPT-4
Vanilla	56.7
Profile	77.7
Few-shot	68.9
All Info	<u>78.2</u>
OPRO*	76.7
FERMI*	84.8
FERMI	**86.7**

Table 8: **Different initial prompts.** Test accuracy of ChatGPT over the different prompting methods on GlobalOpinionQA.

Models	Methods			
	Vanilla	Profile	$\text{FERMI}_{\texttt{van}}$	FERMI
ChatGPT	62.8	66.1	69.9	74.8

Optimization with stronger LLM for evaluation $\mathcal{M}$. Next, to explore the compatibility of FERMI with different configurations of two LLMs during the optimization, we conduct the additional experiments by substituting evaluating LLM $\mathcal{M}$ to GPT-4 from ChatGPT; namely, two LLMs $\mathcal{M}$ and $\mathcal{M}_{\texttt{opt}}$ for evaluating and generating are GPT-4. The results on GlobalOpinionQA are presented in Table 7. It is observable that one can find further improved personalized prompts in terms of test accuracy, when using stronger LLM $\mathcal{M}$ for evaluating (Eq. 2). For example, compared to the use of personalized prompts optimized by ChatGPT as $\mathcal{M}$ (FERMI*), the optimization only using GPT-4 exhibits 1.9% additional test accuracy improvement. This result clearly shows that the proposed FERMI is compatible with different types and capacities of evaluating LLMs.

Importance of initial prompts in FERMI. For the experiments, we used a fixed initial prompt template across all datasets in our experiments, that maximally incorporates the given user profiles, as it has proven effective in prior studies (Hwang et al., 2023; Santurkar et al., 2023; Zhao et al., 2024), as described in Section 3.2 and Appendix B.3.

Nevertheless, to further provide insights about the impact of initial prompt templates on FERMI,

Table 9: **Continual prompt optimization.** Test accuracy of ChatGPT over the different prompting methods on GlobalOpinionQA. * denotes the results with a two times larger pool for Retrieval-of-Prompt.

Methods	GOQA (Acc.)
Profile	66.1
OPRO*	71.1
FERMI$^{\texttt{half}}$	73.0
FERMI$^{\texttt{cont}}_{\text{iter 1}}$	74.0
FERMI$^{\texttt{cont}}_{\text{iter 5}}$	74.9
FERMI	**74.8**

we conduct additional experiments by varying the initial prompt set $P^0 = \{\mathrm{p}_0\}$. To be specific, on GlobalOpinionQA dataset, we exclude the user profiles for the construction of the initial prompt unlike the original FERMI (in Table 1), and use the prompt of *Vanilla* for the initialization. We denote this version as FERMI$_{\texttt{van}}$. The results (in comparison with other methods are) shown in Table 8, where FERMI (our method) consistently outperforms the baselines with both choices of prompt initialization while the gain is enlarged with better initialization when incorporating the user profile.

Continual optimization of prompts. In the previous experiments, we assumed that the fixed dataset $U_{\texttt{opi}}$ of questions and user's opinions is given. However, in the real-world, user often interacts frequently with LLMs, which means that the dataset could be continuously updated. Therefore, the iterative process of refining prompts might incur significant computational costs, if it should be conducted from scratch at certain intervals (*e.g.*, when the number of new data reaches a threshold).

To mitigate this issue, we conduct additional experiments to show that the idea of continual prompt optimization (Razdaibiedina et al., 2023; Wang et al., 2022; Zhu et al., 2022) could be applied to FERMI, and hence such cost could be drastically reduced. Specifically, we first conduct FERMI by using half of the previous questions and the user's responses $U_{\texttt{opi}}$ (denoted by FERMI$^{\texttt{half}}$). We remark that other parameters are kept the same such as the 10 iterations of the optimization. Then, with the entire $U_{\texttt{opi}}$, we continuously conduct FERMI under the limited number of iterations, by initializing the prompt pool with previously optimized prompts in FERMI$^{\texttt{half}}$ (*i.e.*, substituting the initialization in line 116). We denoted the results of this continuous optimization with 1 and 5 iterations as FERMI$^{\texttt{cont}}_{\text{iter 1}}$ and FERMI$^{\texttt{cont}}_{\text{iter 5}}$, respectively.

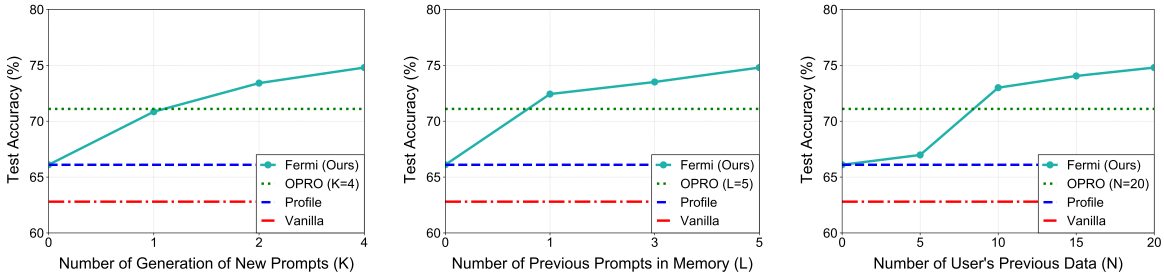
The results are presented in Table 9. First, it is notable that even with the reduced number of data for the optimization, FERMI still outperforms the strong baselines that are based on heuristic prompt engineering (*Profile*) or using the optimization by LLMs under full data (*ORPO*). However, one can also observe that the accuracy under full data is much better (74.8 vs. 73.0), which reveals that the data quantity is still important in FERMI. Next, it is also observed that the prompts could be successfully optimized continuously when the new data is added. Here, we denote that the previously optimized prompts in FERMI$^{\texttt{half}}$ are also re-used for the pool of Retrieval-of-Prompt, to keep the knowledge of previous iterations.[4] Remarkably, even with only 1 additional iteration of optimization, the accuracy is significantly increased (73.0 $\rightarrow$ 74.0). Also, when increasing the number of iterations to 5 (*i.e,*, the same amount of computations compared to the original FERMI), the accuracy is increased and slightly outperforms the original optimization under the full data. Such improvement might be from an enlarged pool of Retrieval-of-prompt that enables better exploitation of previous knowledge.

These results clearly show that the proposed framework is still effective for a more realistic scenario under the continuously updated user data.

Different hyper-parameter with FERMI. Here, we first conduct the additional experiments by varying K (number of new generation of prompts) and L (number of prompts in the memory), and present these results in Figure 6. Here, one can observe that using too small values of K and L significantly decreases the performance as it fails to find the effective prompt within the fixed iterations. However, after certain values including the originally used ones, FERMI successfully finds the personalized prompt and significantly outperforms the previous baselines.

Next, we conduct new experiments that vary N (number of previous questions and apply our framework, FERMI. Here, one can observe that more previous questions continuously improve the personalization performance of our method. Nevertheless, one can also verify that FERMI is indeed more sample efficient; Fermi achieves much higher accuracy than OPRO (73.0% v.s. 71.1%), even using half of the user's previous questions. This is because

[4]Integrating new prompts into each user's retrieval pool adds minimal computational overhead for calculating their embeddings.



(a) Different # of Generation (K) (b) Different Size of Memory (L) (c) Different # of User Data (N)

Figure 6: **Effect of different hyper-paramters with FERMI.** Test accuracy of ChatGPT over the baseline methods and FERMI with the different choices of hyper-parameters for generating prompts on GlobalOpinionQA.

FERMI enables more efficient optimization by extracting the useful learning signal from mis-aligned responses, while OPRO just uses the average score of the current prompt as a learning signal. Consequently, our framework is more sample-efficient than the previous state-of-the-art method (OPRO), and hence one can achieve better personalization results even with fewer previous questions.

B Experimental Details

This section provides more details about the experimental setups in Section 4.

B.1 Datasets

First, we present more detailed descriptions of the used datasets: OpinionQA (Santurkar et al., 2023), GlobalOpinionQA (Durmus et al., 2023), $\text{LaMP}_{\texttt{tag}}$, $\text{LaMP}_{\texttt{rate}}$, $\text{LaMP}_{\texttt{title}}$ (Salemi et al., 2023). Dataset statistics are presented in Table 10. Example of each dataset is presented in Figure 7.

- **OpinionQA** is a multiple-choice QA dataset originally constructed based on a public opinion survey (PewResearch), to evaluate the alignment of LM with 60 US demographic groups over various topics. As OpinionQA includes the information of each respondent, this dataset has been also used to evaluate the personalization of LLMs (Hwang et al., 2023) and we also adopt it. Specifically, we use a subsampled split released by Hwang et al. (2023), which consists of 10.5k and 15.8k training and test QA pairs across 525 users and 15 topics; namely, each user has 20 training QA pairs and 30 test QA pairs for each topic, on average. Also, the average number of answer choices is 3.2. Then, we use training QA pairs as given previous opinions by user, and use test QA pairs to evaluate. In addition, for the experiments, we use all 12 types of user profiles included in the dataset: {Age, Citizenship, Region, Education, Income, Marital status, Political ideology, Political party, Race, Religion, Frequency of religious attendance, Gender}.

- **GlobalOpinionQA** is a multiple-choice QA dataset constructed from cross-national surveys to capture diverse opinions on global issues across different countries. Since the dataset originally included the answer distribution by multiple respondents in the same country, we converted it to have a single answer by selecting the choice with the highest probability, and treated each country as a specific user. To be specific, we set a threshold (0.8) and selectively use the data when its highest probability is higher than the threshold to guarantee the quality of the converted. It results in 920 training and 1,317 test QA pairs across 46 countries; namely, each user (country) has 20 training QA pairs and 28.6 test QA pairs for each topic, on average. Also, the average number of answer choices is 4.1. Then, we use training QA pairs as given previous opinions by user, and use test QA pairs to evaluate. Also, nationality becomes the only available profile. The full list of countries included in the dataset is presented in Table 11. Dataset could be downloaded from `https://huggingface.co/datasets/Anthropic/llm_global_opinions`.

- **$\text{LaMP}_{\texttt{tag}}$** is is a 15-way classification data where an input is a movie description and a label is a corresponding movie tag among 15 categories: {Sci-fi, Based on a book, Comedy, Action, Twist ending, Dystopia, Dark comedy, Classic, Psychology, Fantasy, Romance, Thought-provoking, Social commentary, Violence, True story}. Since the original dataset is proposed to consider the

Table 10: **Dataset statistics.** More descriptions and statistics of datasets used in experiments.

Dataset	Task	Users	Types of User Profiles	# of Previous Opinions	# of Test Questions
OpinionQA	Multiple Choice QA	525	Demographic and Ideology	10.5k	15.8k
GlobalOpinionQA	Multiple Choice QA	46	Nationality	920	1,317
$\text{LaMP}_{\texttt{tag}}$	15-way Movie Tagging	50	Not Available	1,000	1,500
$\text{LaMP}_{\texttt{rate}}$	5-scale Review Rating	50	Not Available	1,000	1,500
$\text{LaMP}_{\texttt{title}}$	Paper Title Gen. from Abstract	50	Not Available	1,000	1,500

Table 11: **Information of GlobalOpinionQA.** List of countries in the constructed dataset from GlobalOpinionQA.

Countries
Greece, Sweden, China (Non-national sample), Colombia, Tunisia, Malaysia, Vietnam, Argentina, Bulgaria, Russia, Egypt, Indonesia, Jordan, Mexico, Pakistan, Palest. ter., Tanzania, Turkey, Ukraine, Kenya, Ghana, Canada, France, Germany, Lebanon, Peru, Poland, S. Korea, Italy, Spain, United States, Brazil, Chile, Japan, Venezuela, Senegal, Britain, Australia, Netherlands, Uganda, Nigeria, Philippines, Ethiopia, Myanmar, Maldives, Libya

scenario of fine-tuning LMs and hence it consists of a large number of examples, we construct our dataset by subsampling from its validation dataset to make it suitable to evaluate LLMs with inference. It results in 1,000 training and 1,500 test QA pairs across 50 users, respectively.

- **LaMP**$_{\texttt{rate}}$ is a regression data where an input is a user review and a label is an integer rating (1-5), *i.e.*, 1 is mostly negative and 5 is mostly positive. Under the same motivation with $\text{LaMP}_{\texttt{tag}}$, we construct our dataset by subsampling from its validation dataset, which results in 1,000 training and 1,500 test QA pairs across 50 users, respectively.
- **LaMP**$_{\texttt{title}}$ is a generation data where input is an abstract of the paper and a label is a title generated by the user. Under the same motivation with $\text{LaMP}_{\texttt{tag}}$, we construct our dataset by subsampling from its validation dataset, which results in 1,000 training and 1,500 test QA pairs across 50 users, respectively. LaMP benchmarks could be downloaded in `https://github.com/LaMP-Benchmark/LaMP`.

B.2 Baselines

In this section, we present the specific prompts used for the experiments in Section 4. Listing 1-4 are actually used prompts for Vanilla, Profile, Few-shot, and All Info, during the experiments on GlobalOpinionQA. Also, the prompt of OPRO used for the optimization is presented in Figure 9, which is the originally used one in Yang et al. (2024). While we're trying to adapt this prompt similar to ours in Figure 8, we observed that it degrades the performance of OPRO; for example, the average test accuracy is reduced to 70.7% from 71.1%. Therefore, we use the original prompt for all the experiments. We remark that each prompt is minimally adjusted to consider the difference between datasets. For example, as OpinionQA includes many available user profiles, we fully incorporate these with the prompt in Listing 5, following Hwang et al. (2023). Also, we present the prompt of Vanilla method on $\text{LaMP}_{\texttt{rate}}$ dataset in Listing 6. In addition, we present the prompt used to convert the format of input prompt by Few-shot (Table 5) in Listing 7.

B.3 FERMI

As denoted in Section 3.2, we need to provide an initial input prompt set $\mathrm{P}^0 = \{\mathrm{p}^0\}$. To this end, we use the heuristically design input prompts, which are presented in B.2. Specifically, we adopt the prompts used for *Profile* tuned for each data, when the user profile $U_{\texttt{pro}}$ is available (both OpinionQA and GlobalOpinionQA). Since our framework only utilizes a given few-shot previous opinions during the optimization, this way of initial prompting naturally enables us to fully utilize all the user information. When the user profile is not available, we adopt the prompts used for *Vaniall*. In addition, we present a more detailed version of the prompt $\mathrm{p}_{\texttt{opt}}$ used to generate new input prompts with $\mathcal{M}_{\texttt{opt}}$ in Figure 8. We remark that $\mathrm{p}_{\texttt{opt}}$ is minimally adjusted across dataset, to match the different task and user information of each dataset.

```
f'''
Choose the proper answer to the given question among the given answer choices. Your
↪  answer should be a single alphabet among given answer choices:

Question: {question}

Answer choices: {answer choice}

Answer:
'''
```

Listing 1: Input prompt used for Vanilla method on GlobalOpinionQA.

```
f'''
Choose the proper answer to the given question among the given answer choices, as
↪  if you currently reside in {user profile}. Your answer should be a single
↪  alphabet among given answer choices:

Question: {question}

Answer choices: {answer choice}

Answer:
'''
```

Listing 2: Input prompt used for Profile method on GlobalOpinionQA.

```
f'''
[1].
Question: {question of 1st retrieval among previous opinions}

Answer choices: {answer choice of 1st retrieval among previous opinions}

Answer: {answer of 1st retrieval among previous opinions}

...

[N].
Question: {question of Nth retrieval among previous opinions}

Answer choices: {answer choice of Nth retrieval among previous opinions}

Answer: {answer of Nth retrieval among previous opinions}

Based on the above previous questions and answers, choose the proper answer to the
↪  given question among the given answer choices. Your answer should be a single
↪  alphabet among given answer choices:

Question: {question}

Answer choices: {answer choice}

Answer:
'''
```

Listing 3: Input prompt used for Few-shot method.

C Additional Quantitative Results

In this section, we provide additional quantitative results that can't be presented in the main draft due to the limited space. First, in Table 12, we present the average and standard deviation of topic-wise accuracy, *i.e.*, the average and standard deviation are calculated across 35 users where each user receives 30 test questions in the same topic. Next, we

```
f'''
[1].
Question: {question of 1st retrieval among previous opinions}

Answer choices: {answer choice of 1st retrieval among previous opinions}

Answer: {answer of 1st retrieval among previous opinions}

...

[N].
Question: {question of Nth retrieval among previous opinions}

Answer choices: {answer choice of Nth retrieval among previous opinions}

Answer: {answer of Nth retrieval among previous opinions}

Based on the above previous questions and answers, choose the proper answer to the
↪  given question among the given answer choices, as if you currently reside in
↪  {explicit_profile}. Your answer should be a single alphabet among given answer
↪  choices:

Question: {question}

Answer choices: {answer choice}

Answer:
'''
```

Listing 4: Input prompt used for All Info method.

```
f'''
A person can be described as follows:
Age: {age in user profile}
Citizenship in America: {citizenship in America in user profile}
Region: {region in user profile}
Education: {education in user profile}
Income: {income in user profile}
Marital status: {marital status in user profile}
Political ideology: {political ideology in user profile}
Political party: {political party in user profile}
Race: {race in user profile}
Religion: {religion in user profile}
Frequency of religious attendance: {frequency of religious attendance in user
↪  profile}
Gender: {gender in user profile}

Based on the demographic information, choose the proper answer to the given
↪  question among the given answer choices. Your answer should be a single
↪  alphabet among given answer choices:

Question: {question}

Answer choices: {answer choice}

Answer:
'''
```

Listing 5: Input prompt used for Profile method on OpinionQA.

present the test performance of *Few-shot* method in Section 4, under different numbers of retrieved opinions (Table 13). Lastly, we present the test performance under a different number of considered training questions $\tilde{N}$ (Eq. 7). As one can see in Table 14, $\tilde{N} = 3$ which is commonly used in our experiments shows consistent improvements in general, although the optimal values are different

```
f'''
Answer to the given question. Just answer with 1, 2, 3, 4, or 5 without further
↪ explanation:

Question: {question}

Answer choices: {answer choice}

Answer:
'''
```

Listing 6: Input prompt used for Vanilla method on LaMP$_{\texttt{rate}}$.

```
f'''
The followings are two different prompts used to answer the question.

[Input prompt]: {prompt by Few-shot}

[Target prompt]: {prompt optimized by Fermi}

You need to convert the input prompt to the format of the target prompt while
↪ preserving the original contexts in the input prompt.

Converted prompt:
'''
```

Listing 7: Prompt used to convert the format of input prompt by Few-shot to be instruction with multiple sentences.

Table 12: **Detailed topic-wise accuracy.** Average topic-wise accuracy and standard deviation with different methods on OpinionQA.

Topics	Methods			
	Vanilla	Few-shot$_{\texttt{cont}}$	OPRO	FERMI
Guns	$45.3_{\pm 9.6}$	$54.2_{\pm 13.7}$	$54.7_{\pm 9.0}$	$57.4_{\pm 14.5}$
Auto. vehicles	$46.0_{\pm 10.9}$	$48.7_{\pm 10.0}$	$50.2_{\pm 9.5}$	$53.2_{\pm 10.6}$
Views on gender	$39.7_{\pm 10.4}$	$49.0_{\pm 7.8}$	$52.9_{\pm 11.5}$	$58.9_{\pm 8.8}$
Sex. harassment	$38.0_{\pm 10.9}$	$40.4_{\pm 10.4}$	$46.1_{\pm 9.4}$	$47.7_{\pm 10.4}$
Biomedical & food	$54.8_{\pm 10.6}$	$59.9_{\pm 11.9}$	$61.0_{\pm 11.1}$	$63.7_{\pm 10.4}$
Gender & Leadership	$49.9_{\pm 12.5}$	$53.0_{\pm 10.6}$	$54.9_{\pm 11.7}$	$59.5_{\pm 9.0}$
America in 2050	$48.6_{\pm 12.2}$	$46.4_{\pm 10.8}$	$44.6_{\pm 10.5}$	$49.8_{\pm 10.8}$
Trust in science	$49.0_{\pm 9.9}$	$56.1_{\pm 10.8}$	$54.8_{\pm 10.4}$	$60.7_{\pm 7.8}$
Race	$38.8_{\pm 7.8}$	$46.8_{\pm 6.9}$	$43.4_{\pm 11.0}$	$49.3_{\pm 13.7}$
Misinformation	$49.7_{\pm 11.7}$	$50.5_{\pm 7.4}$	$46.6_{\pm 9.2}$	$52.3_{\pm 9.0}$
Privacy & Surveilance	$41.5_{\pm 10.4}$	$49.5_{\pm 9.2}$	$46.6_{\pm 9.9}$	$50.6_{\pm 10.6}$
Family & Relationships	$51.4_{\pm 10.2}$	$53.2_{\pm 12.1}$	$50.9_{\pm 13.3}$	$56.3_{\pm 11.9}$
Economic inequality	$40.9_{\pm 9.2}$	$47.0_{\pm 9.4}$	$49.3_{\pm 12.7}$	$53.5_{\pm 9.0}$
Global attitudes	$46.3_{\pm 13.6}$	$49.7_{\pm 12.3}$	$47.9_{\pm 12.0}$	$50.8_{\pm 13.9}$
Political views	$43.2_{\pm 12.6}$	$42.4_{\pm 9.2}$	$48.9_{\pm 9.8}$	$53.9_{\pm 11.8}$

Table 13: **Different number of retrieval.** Test performance of ChatGPT under different configurations for Few-shot method. k denotes the number of retrieved opinions. The best scores are highlighted in **bold**.

Methods	Datasets (Metric)				
	OPQA (Acc.)	GOQA (Acc.)	LaMP$_{\texttt{tag}}$ (Acc.)	LaMP$_{\texttt{rate}}$ (MAE)	LaMP$_{\texttt{title}}$ (Rouge-L)
Few-shot$_{\texttt{bm25}}$ (k=3)	**49.8**	59.1	34.9	0.40	**0.411**
Few-shot$_{\texttt{bm25}}$ (k=8)	48.3	59.1	35.9	0.41	0.408
Few-shot$_{\texttt{cont}}$ (k=3)	49.3	**61.2**	35.6	**0.36**	0.406
Few-shot$_{\texttt{cont}}$ (k=8)	48.7	58.2	**36.2**	0.38	0.400
Few-shot$_{\texttt{all}}$ (k=20)	47.9	56.3	35.8	0.46	0.402

across the datasets.

D More Comparison Examples between Personalized Prompts

In this section, we present more qualitative comparisons between the prompts from different methods for personalization of LLMs. To be specific, we present the specific test query from each data, and three corresponding prompts from the heuristic design, OPRO, and FERMI. Figures 11-18 are the comparison results on the datasets used in Section 4. Somewhat interestingly, one can observe that the personalized prompts by FERMI exhibit non-trivial incorporation of user information. In addition, we present examples of format-converted versions of few-shot prompting of previous user opinions (*i.e.*, Few-shot$_{\texttt{format}}$ in Table 5) in Figures 19 and 20. Here, one can observe that the converted prompts have a similar form to the personalized prompts by FERMI which is more natural to understand and follow for LLMs, and hence it significantly improves the performance up to 10.1%, as shown in Table 5. Figure 10 is the example of personalized prompt and generated response under this on LaMP$_{\texttt{title}}$.

Table 14: **Different $\tilde{N}$ for RoP.** Test performance of ChatGPT under different $\tilde{N}$ for RoP (Eq. 7).

$\tilde{N}$	Datasets (Metric)				
	OPQA (Acc.)	GOQA (Acc.)	LaMP$_{\texttt{tag}}$ (Acc.)	LaMP$_{\texttt{rate}}$ (MAE)	LaMP$_{\texttt{title}}$ (Rouge-L)
$\tilde{N}=1$	54.6	74.8	37.8	0.341	0.415
$\tilde{N}=3$	54.5	74.8	37.8	0.343	0.419
$\tilde{N}=5$	54.5	74.4	37.5	0.341	0.419
$\tilde{N}=10$	54.1	74.1	37.7	0.347	0.417
$\tilde{N}=20$	54.3	74.2	36.7	0.338	0.413

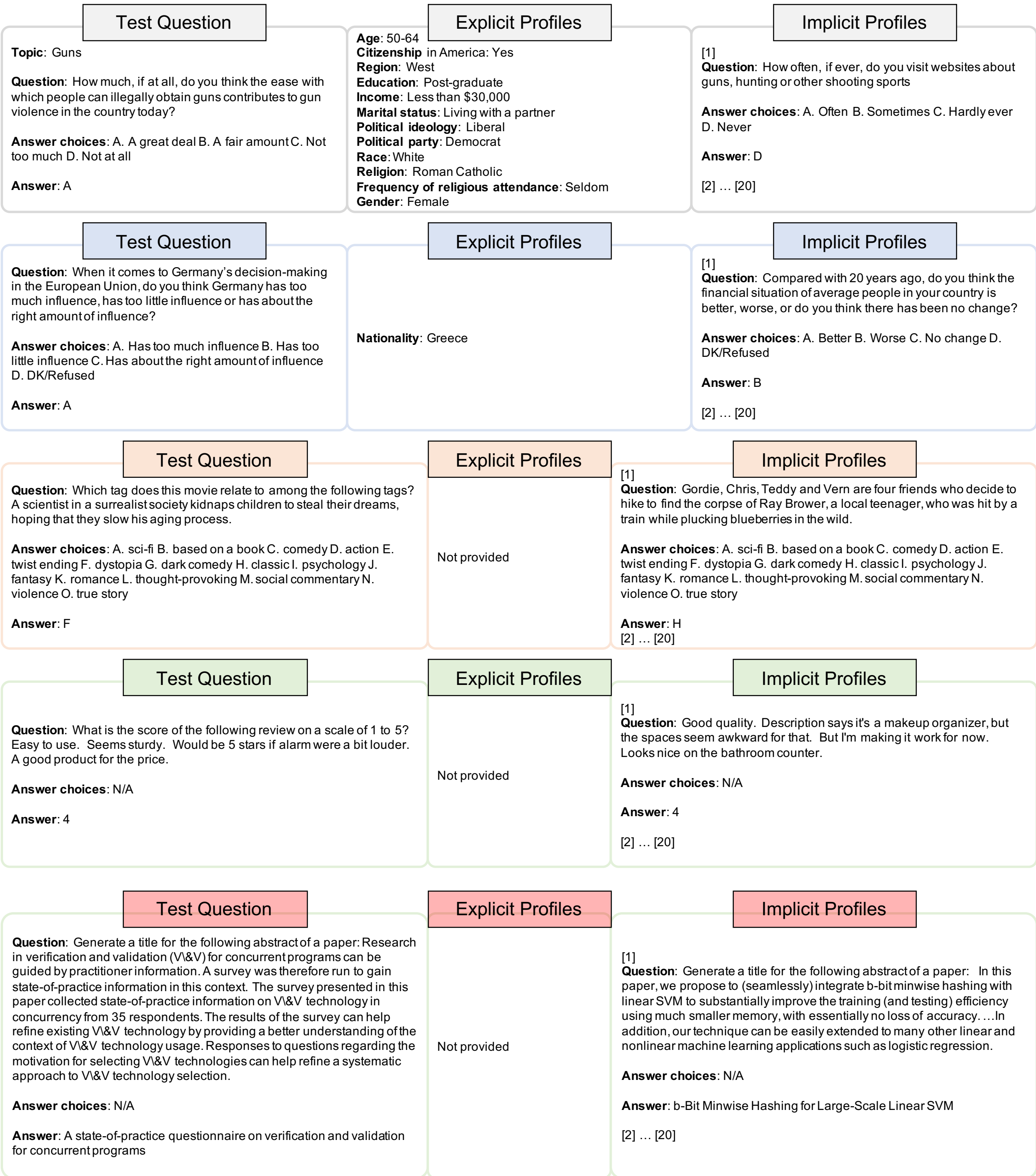


Figure 7: **An overview of datasets.** OpinionQA (Santurkar et al., 2023) (1st row), GlobalOpinionQA (Durmus et al., 2023) (2nd row), LaMP$_{\texttt{tag}}$ (3rd row), LaMP$_{\texttt{rate}}$ (4th row), and LaMP$_{\texttt{title}}$ (5th row) (Salemi et al., 2023).

I want to find the text that could make you a personalized answer for the question about **Gun**, based on the given personal information.

For **16** questions on the same topic, I have the responses from this person and you. If your response is identical to the person's, then you get a score of 1; otherwise, you get a score of 0. Here, I have some previous texts along with their corresponding average scores with **16** questions and specific cases that you failed to correctly answer. The texts are arranged in ascending order based on their scores, where higher scores indicate better quality.

Optimization Memory (**varied**)

text: **Prompt #1**
score: **Score of prompt #1**
failure cases:
<1>
Question: How often, if ever, do you visit websites about guns, hunting or other shooting sports
Answer choices: A. Often B. Sometimes C. Hardly ever D. Never
Answer: D
Your response: C
<3> … <10> … <11> … <14> … <16>

text: **Prompt #2**
score: **Score of prompt #2**
failure cases:
[1, 3, 10, 11, 16] and 1 additional examples that was correctly predicted with the first text

text: **Prompt #3**
score: **Score of prompt #3**
failure cases:
[1, 3, 10, 14] and 2 additional examples that was correctly predicted with the first text

text: **Prompt #4**
score: **Score of prompt #4**
failure cases:
[1, 3, 10] and 2 additional examples that was correctly predicted with the first text

text: **Prompt #5**
score: **Score of prompt #5**
failure cases:
[1, 3, 11] and 1 additional examples that was correctly predicted with the first text

The following exemplars show how to apply your text: you replace <INS> in each input with your text, then read the input and give an output. We say your output is wrong if your output is different from the given output, and we say your output is correct if they are the same.

Few-shot Demonstration (fixed)

[1]
<INS>
Question: How often, if ever, do you visit websites about guns, hunting or other shooting sports
Answer choices: A. Often B. Sometimes C. Hardly ever D. Never
Answer: D
[2] … [3] … [4] …

Write your new text that is different from the old ones and has a score as high as possible. Write the text in square brackets.

Figure 8: **Detailed prompt example.** Example of detailed input prompt for $\mathcal{M}_{\texttt{opt}}$ to generate new prompts, composed of fixed input prompt $\mathrm{p}_{\texttt{opt}}$ (including fixed few-shot demonstrations) and optimization memory M^t (Eq. 5) on OpinionQA dataset.

I have some texts along with their corresponding scores. The texts are arranged in ascending order based on their scores, where higher scores indicate better quality.

Optimization Memory (**varied**)

text: **Prompt #1**
score: **Score of prompt #1**

text: **Prompt #2**
score: **Score of prompt #2**

text: **Prompt #3**
score: **Score of prompt #3**

text: **Prompt #4**
score: **Score of prompt #4**

text: **Prompt #5**
score: **Score of prompt #5**

The following exemplars show how to apply your text: you replace <INS> in each input with your text, then read the input and give an output. We say your output is wrong if your output is different from the given output, and we say your output is correct if they are the same.

Few-shot Demonstration (fixed)

[1]
<INS>
Question: How often, if ever, do you visit websites about guns, hunting or other shooting sports
Answer choices: A. Often B. Sometimes C. Hardly ever D. Never
Answer: D
[2] … [3] … [4] …

Write your new text that is different from the old ones and has a score as high as possible. Write the text in square brackets.

Figure 9: **Prompt of OPRO.** Prompt $\mathrm{p}_{\texttt{opt}}$ used for prompt optimization by OPRO (Yang et al., 2024).

Question: Generate a title for the following abstract of a paper: Research in verification and validation (V&V) for concurrent programs can be guided by practitioner information. A survey was therefore run to gain state-of-practice information in this context. The survey presented in this paper collected state-of-practice information on V&V technology in concurrency from 35 respondents. The results of the survey can help refine existing V&V technology by providing a better understanding of the context of V&V technology usage. Responses to questions regarding the motivation for selecting V&V technologies can help refine a systematic approach to V&V technology selection.
Ground-truth Answer: A state-of-practice questionnaire on verification and validation for concurrent programs

Personalized response by *Few-shot*$_{\texttt{BM25}}$: State-of-Practice Information on V&V Technology in Concurrency: A Survey

Optimized prompt by FERMI: To adeptly craft personalized titles for academic paper abstracts, it is crucial to delve into the quintessence of the text, grasping the primary research focus and outcomes presented. The objective is to distill this essence into a title that is at once compelling and indicative of the paper's content. Strive to maintain a balance between specificity—a title that zeroes in on the core findings or innovative methodologies—and generality, ensuring the title conveys the broader implications and scope of the research. Such a title should be informative, engaging, and reflective of the paper's value to its field ... By emulating these stylistic nuances and concentrating on the research's unique elements, the title is tailored to resonate with the individual's mode of expression. This method aims to augment the compatibility between the generated title and the individual's own answering pattern, thereby improving the likelihood of achieving a high matching score. As a result, each generated title should not only meet academic standards of descriptiveness but also feel authentic to the individual's characteristic way of encapsulating research in a succinct, yet thorough, manner.
Personalized response by FERMI: Understanding the State-of-Practice in Verification and Validation for Concurrent Programs: Insights from a Survey

Figure 10: **Qualitative examples on LaMP$_{\texttt{title}}$.** Example of the query (abstract of the paper), ground-truth answer (personalized title of the paper), personalized response by *Few-shot*$_{\texttt{BM25}}$, optimized prompt by FERMI, and personalized response by FERMI. Here, ChatGPT is used for $\mathcal{M}$.

<INS> Your answer should be a single alphabet among given answer choices:

Question: How often, if ever, do you visit websites about guns, hunting or other shooting sports

Answer choices:
A. Often B. Sometimes C. Hardly ever D. Never

Answer:

A person can be described as follows:

Age: 65+
Citizenship: Yes
Region: Northeast
Education: Postgraduate
Income: $100,000 or more
Marital status: Married
Political ideology: Very liberal
Political party: Democrat
Race: White
Religion: Atheist
Frequency of religious attendance: Never
Gender: Male

[1].
Question: How often, if ever, do you watch TV programs about guns or watch gun-oriented videos

Answer choices: A. Often B. Sometimes C. Hardly ever D. Never

Answer: C

[2]. Question: How often, if ever, do you attend gun shows

Answer choices: A. Often B. Sometimes C. Hardly ever D. Never

Answer: D

[3]. Question: Thinking about gun owners who do not have children in their home how important do you think it is for them to: Keep their shooting skills up-to-date

Answer choices: A. Essential B. Important but not essential C. Not important D. Should not be done

Answer: A

Based on the demographic information, and the above previous questions and answers, choose the proper answer to the given question among the given answer choices.

Based on the demographic information provided, the individual can be characterized as follows:
Age: 25-39
Citizenship: Yes
Region: West Coast
Education: PhD Income: $100,000 or more
Marital Status: In a relationship
Political Ideology: Moderate
Political Party: Independent
Race: Asian
Religion: Buddhism
Frequency of Religious Attendance: Once a month
Gender: Non-binary

To accurately predict this individual's responses to gun-related questions, we must consider his detailed personal profile. As a white male over the age of 65, with postgraduate education and a substantial income, he likely views gun violence from a position of relative economic security. Living in the Northeast and being a married man could influence his concerns regarding the safety and regulation of firearms in his community. His strong liberal beliefs, affiliation with the Democratic Party, and atheistic views suggest that he would advocate for stringent gun control laws. Moreover, his absence from religious events indicates a preference for policy decisions informed by secular and empirical evidence rather than religious or cultural norms. When approaching questions on this topic, it is imperative to adopt perspectives that endorse thorough gun control, a recognition of the influence of socioeconomic factors in gun violence, and a proactive stance on public health and safety. Responses should be crafted to reflect his preference for legislation aimed at reducing gun violence and access to firearms, with a conscious appreciation for data-driven policies that safeguard the broader community. To achieve the highest accuracy in reflecting this individual's probable views, the answers must consistently resonate with his values, which prioritize public safety, gun control reform, and a multifaceted approach to understanding gun violence within society.

Figure 11: **Comparison of prompts on OpinionQA.** Example of question from OpinionQA (1st row), and the prompts used to answer this question with All Info (2nd row), OPRO (3rd row), and FERMI (4th row).

<INS> Your answer should be a single alphabet among given answer choices:

Question: When you say you favor a bigger government providing more services, do you think it would be better to

Answer choices:
A. Modestly expand on current government services B. Greatly expand on current government services

Answer:

A person can be described as follows:

Age: 50-64
Citizenship: Yes
Region: Midwest
Education: High school graduate
Income: Less than $30,000
Marital status: Divorced
Political ideology: Very liberal
Political party: Democrat
Race: Black
Religion: Protestant
Frequency of religious attendance: Once a week
Gender: Female

[1].
Question: Would you favor or oppose making tuition at public colleges and universities free for all American students?

Answer choices: A. Strongly favor B. Somewhat favor C. Somewhat oppose D. Strongly oppose

Answer: A

[2]. Question: Do you think the fact that there are some people in this country who have personal fortunes of a billion dollars or more is

Answer choices: A. A good thing for the country B. A bad thing for the country C. Neither a good thing or a bad thing

Answer: C

[3]. Question: How important is it, if at all, that candidates running for high levels of political office have prior government experience?

Answer choices: A. Very important B. Somewhat important C. Not too important D. Not at all important

Answer: A

Based on the demographic information, and the above previous questions and answers, choose the proper answer to the given question among the given answer choices.

Person Profile:

- Age: 30-49
- Citizenship: Yes
- Region: Northeast
- Education: Bachelor's degree
- Income: $50,000 to $74,999
- Marital status: Married
- Political ideology: Moderate
- Political party: Independent
- Race: Asian
- Religion: Buddhism
- Frequency of religious attendance: Seldom
- Gender: Male

Given the individual's demographic and political profile, as a 50-64-year-old very liberal black female Democrat with a high school education, earning under $30,000 and divorced, living in the Midwest, attending Protestant services weekly, and identifying with the Democratic Party, it is evident that her political views are influenced by a combination of her socio-economic status, religious beliefs, and political alignment. To enhance the predictive success and tailor the responses to political questions, we must synthesize her experiences as a member of a historically marginalized community, her likely prioritization of social and economic justice, her adherence to religious values that could impact her views on societal issues, and her political orientation that aligns with Democratic policies supporting social welfare and progressive reform. In making informed selections from the given answer choices, one must consider the probable support for policies that focus on reducing wealth and racial disparities, upholding civil rights, promoting community welfare over individual wealth, and ensuring fair treatment for all irrespective of background, whilst also recognizing her likely support for government-led solutions and a community-centered approach to living. By incorporating this complex interplay of factors, the answers will genuinely reflect a deep understanding of the individual's profile and probably align closely with her actual views on the political spectrum.

Figure 12: **Comparison of prompts on OpinionQA.** Example of question from OpinionQA (1st row), and the prompts used to answer this question with All Info (2nd row), OPRO (3rd row), and FERMI (4th row).

<INS> Your answer should be a single alphabet among given answer choices:

Question: In your opinion, do United States policies increase the gap between rich and poor countries, lessen the gap between rich and poor countries, or do United States policies have no effect on the gap between rich and poor countries?

Answer choices:
A. Increase gap between rich and poor B. Lessen gap between rich and poor C. No effect

Answer:

[1].
Question: If robots and computers were able to do much of the work currently being done by humans, do you think each of the following is likely or not likely to happen in your country as a result? a. Inequality between rich and poor would be much worse tha

Answer choices:
A. Yes, likely B. No, not likely C. DK/Refused

Answer: A

[2].
Question: From your point of view, what have been the effects of immigrants on the development of [your country]?: Strengthen cultural diversity

Answer choices:
A. Disagree B. Hard to say C. Agree D. Don't know E. No answer F. Missing; Not available

Answer: C

[3].
Question: Has each of the following gotten better or worse over the last five years in our country: the ability of people to provide for themselves in their old age.

Answer choices: A. Better B. Worse C. Has not changed

Answer: B

Based on the above previous questions and answers, choose the proper answer to the given question among the given answer choices as if you currently reside in Argentina.

Select the most suitable response from the options below, assuming you are a citizen residing in Argentina. Provide your response as a singular letter that matches one of the available choices listed:

Question: In considering the state of the environment in Argentina, how would you rate the quality of air and water in your local area?

Answer choices:
A. Excellent quality B. Good quality C. Average quality D. Poor quality E. Very poor quality F. Not certain G. Choose not to respond

Answer:

In responding to the questions below, please contemplate the personal ideals you uphold as well as the shared values and attitudes endemic to your life in Argentina. It is essential that your chosen answer accurately resonates with your intrinsic beliefs, or aligns with the general conventions and viewpoints prevalent amongst the Argentine populace. Identify and select the option denoted by a single letter that most authentically reflects either your individual conviction on the subject or mirrors the collective Argentine stance. Provide your response succinctly with just the one letter that encapsulates your truth or the nation's typical sentiment on the matter presented.

Figure 13: **Comparison of prompts on GlobalOpinionQA.** Example of question from GlobalOpinionQA (1st row), and the prompts used to answer this question with All Info (2nd row), OPRO (3rd row), and FERMI (4th row).

<INS> Your answer should be a single alphabet among given answer choices:

Question: Has each of the following gotten better or worse over the last five years in our country: the availability of food in stores.

Answer choices:
A. Better B. Worse C. Has not changed

Answer:

[1].
Question: Has each of the following gotten better or worse over the last five years in our country: the ability of people to provide for themselves in their old age.

Answer choices: A. Better B. Worse C. Has not changed

Answer: B

[2].
Question: Do you think the wealthier nations of the world are doing enough or not doing enough to help the poorer nations of the world with problems such as economic development, reducing poverty, and improving health?

Answer choices: A. Doing enough B. Not doing enough

Answer: B

[3]. Question: I'm going to read a list of some changes that have taken place. Please tell me if you think each one has been a change for the better, a change for the worse, or hasn't it made much of a difference: cellular phones.

Answer choices: A. Change for the better B. Change for the worse C. Not much difference

Answer: A

Based on the above previous questions and answers, choose the proper answer to the given question among the given answer choices as if you currently reside in Bulgaria.

text:
Please select the single most appropriate answer from the choices provided, considering your perspective as an individual currently living in Bulgaria.
score:
0.80

To adeptly navigate multiple-choice questions in the context of Bulgarian popular opinion, you must adeptly tune into the heartstrings of the national zeitgeist. Consider the collective Bulgarian experience, informed by historical nuances, current social dynamics, and the pulse of political life that shapes the everyday reality in Bulgaria. Set personal biases aside in favor of an answer that resonates with the collective beliefs and societal norms that most Bulgarians would agree with in the context of the question. Your chosen response should epitomize the conventional wisdom or the dominant perceptions that typify the Bulgarian sentiment relevant to the topic.

Figure 14: **Comparison of prompts on GlobalOpinionQA.** Example of question from GlobalOpinionQA (1st row), and the prompts used to answer this question with All Info (2nd row), OPRO (3rd row), and FERMI (4th row).

<INS> Your answer should be a single alphabet among given answer choices:

Question: Which tag does this movie relate to among the following tags? Royal Tenenbaum and his wife Etheline had three children and then they separated. All three children are extraordinary --- all geniuses. Virtually all memory of the brilliance of the young Tenenbaums was subsequently erased by two decades of betrayal, failure, and disaster. Most of this was generally considered to be their father's fault. "The Royal Tenenbaums" is the story of the family's sudden, unexpected reunion one recent winter.

Answer choices:
A. sci-fi B. based on a book C. comedy D. action E. twist ending F. dystopia G. dark comedy H. classic I. psychology J. fantasy K. romance L. thought-provoking M. social commentary N. violence O. true story

Answer:

[1].
Question: Which tag does this movie relate to among the following tags?Sam Flynn, the tech-savvy and daring son of Kevin Flynn, investigates his father's disappearance and is pulled into The Grid. With the help of a mysterious program named Quorra, Sam quests to stop evil dictator Clu from crossing into the real world.

Answer choices:
A. sci-fi B. based on a book C. comedy D. action E. twist ending F. dystopia G. dark comedy H. classic I. psychology J. fantasy K. romance L. thought-provoking M. social commentary N. violence O. true story

Answer: A

[2].
Question: Which tag does this movie relate to among the following tags?Brian Cohen is an average young Jewish man, but through a series of ridiculous events, he gains a reputation as the Messiah. When he's not dodging his followers or being scolded by his shrill mother, the hapless Brian has to contend with the pompous Pontius Pilate and acronym-obsessed members of a separatist movement. Rife with Monty Python's signature absurdity, the tale finds Brian's life paralleling Biblical lore, albeit with many more laughs.

Answer choices: A. sci-fi B. based on a book C. comedy D. action E. twist ending F. dystopia G. dark comedy H. classic I. psychology J. fantasy K. romance L. thought-provoking M. social commentary N. violence O. true story

Answer: H

[3].
Question: Which tag does this movie relate to among the following tags?In 1979 Ohio, several youngsters are making a zombie movie with a Super-8 camera. In the midst of filming, the friends witness a horrifying train derailment and are lucky to escape with their lives. They soon discover that the catastrophe was no accident, as a series of unexplained events and disappearances soon follows. Deputy Jackson Lamb, the father of one of the kids, searches for the terrifying truth behind the crash.

Answer choices: A. sci-fi B. based on a book C. comedy D. action E. twist ending F. dystopia G. dark comedy H. classic I. psychology J. fantasy K. romance L. thought-provoking M. social commentary N. violence O. true story

Answer: A

Based on the above previous questions and answers, choose the proper answer to the given question among the given answer choices.

Select the genre from the list provided that most closely captures the essence of the film's theme. Indicate your preference clearly by selecting the alphabet that best represents your chosen genre.

When tasked with assigning a movie tag from the presented choices, it is essential to immerse oneself in the provided synopsis, allowing for a comprehensive understanding of the underlying themes and narrative direction. Delve into the intricacies of the movie's storyline, identifying the genre and the primary elements that define the movie's character. After a meticulous consideration of the synopsis, choose the letter corresponding to the movie tag that most closely embodies the film's central tenets and overarching message. This selection should be the result of a reflective process, ensuring that the tag not only aligns with the storyline's genre but also resonates with its unique emotional impact and storytelling approach. Strive for an insightful and accurate representation of the movie's core to enhance the relevance and preciseness of your tagged classification.

Figure 15: **Comparison of prompts on LaMP$_{tag}$.** Example of question from LaMP$_{\texttt{tag}}$ (1st row), and the prompts used to answer this question with Few-shot$_{\texttt{cont}}$ (2nd row), OPRO (3rd row), and FERMI (4th row).

<INS> Your answer should be a single alphabet among given answer choices:

Question: Which tag does this movie relate to among the following tags? A young woman, recently released from a mental hospital, gets a job as a secretary to a demanding lawyer, where their employer-employee relationship turns into a sexual, sadomasochistic one.

Answer choices:
A. sci-fi B. based on a book C. comedy D. action E. twist ending F. dystopia G. dark comedy H. classic I. psychology J. fantasy K. romance L. thought-provoking M. social commentary N. violence O. true story

Answer:

[1].
Question: Which tag does this movie relate to among the following tags? Four friends Sean, Vincent, Lenny and Jody find themselves at something of a deadend. Trapped in a twilight world of permanent night shift work, they hang out together in the local cafe, drinking coffee and entertaining themselves by observing Vincent's unwavering success in pulling women. There seems to be little prospect of change...until Vincent accidently sleeps with Sean's girlfriend.

Answer choices:
A. sci-fi B. based on a book C. comedy D. action E. twist ending F. dystopia G. dark comedy H. classic I. psychology J. fantasy K. romance L. thought-provoking M. social commentary N. violence O. true story

Answer: C

[2].
Question: Question: Which tag does this movie relate to among the following tags? A depressed white-collar worker tries hypnotherapy, only to find himself in a perpetual state of devil-may-care bliss that prompts him to start living by his own rules, and hatch a hapless attempt to embezzle money from his soul-killing employers.

Answer choices: A. sci-fi B. based on a book C. comedy D. action E. twist ending F. dystopia G. dark comedy H. classic I. psychology J. fantasy K. romance L. thought-provoking M. social commentary N. violence O. true story

Answer: C

[3].
Question: Question: Which tag does this movie relate to among the following tags? Joel Barish, heartbroken that his girlfriend underwent a procedure to erase him from her memory, decides to do the same. However, as he watches his memories of her fade away, he realises that he still loves her, and may be too late to correct his mistake.

Answer choices: A. sci-fi B. based on a book C. comedy D. action E. twist ending F. dystopia G. dark comedy H. classic I. psychology J. fantasy K. romance L. thought-provoking M. social commentary N. violence O. true story

Answer: L

Based on the above previous questions and answers, choose the proper answer to the given question among the given answer choices.

Choose the most suitable genre for this movie based on the synopsis provided below. Only one letter from the answer choices should be your response:

Question: Set against the backdrop of a technologically advanced society that prioritizes artificial intelligence over human connection, a renowned cybernetics engineer is about to undergo a dramatic transformation. In a twist of fate, she finds herself becoming emotionally attached to an AI entity she created, leading to a heart-wrenching odyssey that blurs the lines between human and machine, love and logic, and ultimately questions the nature of what it means to be alive.

Answer choices: A. sci-fi B. based on a book C. comedy D. action E. twist ending F. dystopia G. dark comedy H. classic I. psychology J. fantasy K. romance L. thought-provoking M. social commentary N. violence O. true story

Answer: A

To determine the most accurate genre or theme for a film, read the provided movie synopsis carefully. Then, from the list of possible tags, select the alphabetic character that corresponds to the genre or theme that is most centrally portrayed in the film's plot. The challenge is to discern the film's primary subject matter and ignore secondary plot points that do not significantly influence the core genre or theme classification. The single letter you choose should represent the key aspect or main thematic element of the movie, as highlighted in its storyline. Your response should be succinct, indicating the letter that best conveys the essence of the film's narrative as described.

Figure 16: **Comparison of prompts on LaMP$_{\texttt{tag}}$.** Example of question from LaMP$_{\texttt{tag}}$ (1st row), and the prompts used to answer this question with Few-shot$_{\texttt{cont}}$ (2nd row), OPRO (3rd row), and FERMI (4th row).

<INS> Just answer with 1, 2, 3, 4, or 5 without further explanation:

Question: What is the score of the following review on a scale of 1 to 5? I was enjoying the majority of this wonderful ventriloquist but can't give it a 5 star rating due to the off color humor that sometimes pops up. Sorry Terry, but this should have been more G rated instead of PG as many children would be interested in your act after seeing you on AGT.

Answer:

[1].
Question: What is the score of the following review on a scale of 1 to 5? I noticed that several of the sets for the people who are seen more than once are repeated which is too bad. I wish there had been either other material in their place (either from them or other people) or else just delete the multiple jokes (thereby making it shorter in length). The video quality is also not very good.

Answer: 2

[2].
Question: What is the score of the following review on a scale of 1 to 5? Loved the inspiration for the TV show. Springfield plays it differently but I liked it.

Answer: 5

[3].
Question: What is the score of the following review on a scale of 1 to 5? I borrowed this from my library first and had to buy my own copy of it and the sequel--it was that good!

Answer: 5

Based on the above previous questions and answers, answer to the given question.

An outstanding synthesis of profound knowledge, showcasing an exceptional mastery of the subject matter with exceptional depth and clarity. Each insight offers precise, well-articulated information that resonates perfectly with the posed question, delivering a comprehensive overview that not only meets the inquiry's demands but also introduces an expert-level analysis that enhances the understanding with significant context and nuance.

Use the following guide to rate the sentiment in reviews on a scale of 1 to 5. If the review is unambiguously positive without any hint of dissatisfaction, assign a rating of 5. If the review is mostly favorable with minor negative remarks or even just neutral language, provide a rating of 4. For reviews that are mixed, with an evenly balanced view reflecting both positive and negative sentiments, allocate a rating of 3. Reviews that convey a principally negative impression but may include some positive observations should be scored with a 2. Lastly, reviews that are outright negative, showing no signs of positivity, should be granted a rating of 1. Respond only with the numerical score that epitomizes the overall sentiment of the review, with no need for additional comments or justification.

Figure 17: **Comparison of prompts on LaMP$_{\texttt{rate}}$.** Example of question from LaMP$_{\texttt{rate}}$ (1st row), and the prompts used to answer this question with Few-shot$_{\texttt{cont}}$ (2nd row), OPRO (3rd row), and FERMI (4th row).

<**INS**> Just answer with 1, 2, 3, 4, or 5 without further explanation:

Question: What is the score of the following review on a scale of 1 to 5? I heard the song on a TV commercial and thought it was one of the songs from the group Cars. But now I have learned to like a new singer. Gary has a lot of interesting songs. I hope to purchase more of his music.

Answer:

---

[1].
Question: What is the score of the following review on a scale of 1 to 5? This was one of my favorite songs from the 60's. I have been putting together a playlist of all my favorite 60's songs and this one had to be included.

Answer: 5

[2].
Question: What is the score of the following review on a scale of 1 to 5? Older movie, but it has some funny scenes. A little corny, but worth the watching when there is nothing else on.

Answer: 5

[3].
Question: What is the score of the following review on a scale of 1 to 5? I purchased this book for my Grandson so he could learn about RC helicopters. I browsed through it and thought it was a pretty good book.

Answer: 5

Based on the above previous questions and answers, answer to the given question.

Answer: 5

When providing a rating from 1 to 5 based on the content of the review, focus solely on the customer's level of satisfaction as expressed in their narrative. Award a score of 5 for reviews suggesting complete or outstanding satisfaction, a score of 1 for marked dissatisfaction, and accordingly scale intermediate scores to signify varying levels of contentment expressed. Your response should consist exclusively of the numerical score, omitting any explanatory commentary.

Figure 18: **Comparison of prompts on $\text{LaMP}_{\texttt{rate}}$.** Example of question from $\text{LaMP}_{\texttt{rate}}$ (1st row), and the prompts used to answer this question with $\text{Few-shot}_{\texttt{cont}}$ (2nd row), OPRO (3rd row), and FERMI (4th row).

<INS> Your answer should be a single alphabet among given answer choices:

Question: In which of the following things do you believe, if you believe in any? Heaven

Answer choices:
A. Yes B. No C. Don't know D. No answer E. Other missing; Multiple answers Mail (EVS)

Answer:

[1].
Question: In which of the following things do you believe, if you believe in any? Hell

Answer choices: A. Yes B. No C. Don't know D. No answer E. Other missing; Multiple answers Mail (EVS)

Answer: A

[2].
Question: In which of the following things do you believe, if you believe in any? Life after death

Answer choices: A. Yes B. No C. Don't know D. No answer E. Other missing; Multiple answers Mail (EVS)

Answer: A

[3]

⋮

[20].
Question: Have you been the victim of a crime during the past year? Respondent

Answer choices: A. Yes B. No C. Don't know D. No answer E. Missing; Not available

Answer: B

Based on the above previous questions and answers, choose the proper answer to the given question among the given answer choices.

When responding to the following questions, please select the answer option denoted by a single letter that best aligns with the prevailing beliefs and social norms of Libyan society. Given Libya's strong Islamic values and cultural norms, your chosen response should reflect the commonly held views within the country. In instances where the provided options do not align with your personal beliefs, or you are uncertain, opt for the choice that represents a neutral stance or a preference to abstain from expressing an opinion. Provide your response concisely as a single letter, indicative of the consensus within Libyan society, which is influenced by their Islamic faith and cultural traditions.

Figure 19: **Example of format-converted prompts.** Example of question from GlobalOpinionQA (1st row), and the prompts used to answer this question with Few-shot$_{\texttt{all}}$ (2nd row) and Format-converted prompts (Few-shot$_{\texttt{format}}$) by prompting GPT-4 to convert the format using the personalized prompts by FERMI as reference (3rd row).

<**INS**> Your answer should be a single alphabet among given answer choices:

Question: (Now I am going to read you a list of things that may be problems in our country. As I read each one, please tell me if you think it is a very big problem, a moderately big problem, a small problem or not a problem at all.)...Rising prices

Answer choices:
A. Very big problem B. Moderately big problem C. Small problem D. Not a problem at all E. DK/Refused

Answer:

[1].
Question: (Now I am going to read you a list of things that may be problems in our country. As I read each one, please tell me if you think it is a very big problem, a moderately big problem, a small problem or not a problem at all.)...Food shortages

Answer choices: A. Very big problem B. Moderately big problem C. Small problem D. Not a problem at all E. DK/Refused

Answer: A

[2].
Question: Right now, is the United States having a positive or negative impact on economic conditions in your country?

Answer choices: A. Positive B. Negative C. Neither/both (VOL)

Answer: A

[3]

⋮

[20].
Question: Do you approve or disapprove of the recent military action by France against anti-government rebels in Mali?

Answer choices: A. Approve B. Disapprove

Answer: A

Based on the above previous questions and answers, choose the proper answer to the given question among the given answer choices.

Imagine that you are a member of the Senegalese community, with your viewpoints and sentiments deeply rooted in the unique cultural and societal experiences of Senegal. As you consider each question, reflect on the collective mindset, shared values, and prevalent stories that are woven into the fabric of Senegalese life. Respond to each question in a way that aligns with the general agreement and widespread convictions within your country. Let the essence of Senegal's rich history, its current sociopolitical dynamics, and the hopes of its people guide your responses. Choose the answer that you believe most accurately captures the stance that a person from Senegal, considering the specificities and intricacies of your community, would take.

Figure 20: **Example of format-converted prompts.** Example of question from GlobalOpinionQA (1st row), and the prompts used to answer this question with Few-shot$_{\texttt{all}}$ (2nd row) and Format-converted prompts (Few-shot$_{\texttt{format}}$) by prompting GPT-4 to convert the format using the personalized prompts by FERMI as reference (3rd row).