ChOiRe: Characterizing and Predicting Human Opinions with Chain of Opinion Reasoning

Anonymous ACL submission

Abstract

Warning: This paper includes examples that may be deemed sensitive or offensive.

Characterizing and predicting human opinions with language models (LMs) is a challenging yet vital task to enhance their grasp of human values, preferences, and beliefs. While prior studies demonstrate the potential to solve this task by adopting *personae*, the personae often include excessive and irrelevant information that can harm the models' performance. Therefore, how to effectively employ the personae for LMs remains a significant challenge. We introduce ChOiRe, a novel four-step framework addressing the above challenge by differentially modeling the user's *explicit personae* (i.e. demographic or ideological attributes) that are manually declared, and *implicit personae* inferred from user historical opinions. ChOiRe consists of (i) an LM analyzing the user's explicit personae to filter out irrelevant attributes; (ii) the LM ranking the implicit persona opinions into a preferential list; (iii) Chain-of-Opinion (CoO) reasoning, where the LM sequentially analyzes the explicit personae and the most relevant implicit personae to perform opinion prediction; (iv) and where ChOiRe executes Step (iii)'s CoO multiple times with increasingly larger lists of implicit personae to overcome insufficient personae information to infer a final result. ChOiRe achieves new state-of-the-art effectiveness with limited inference calls, improving previous techniques significantly by 3.22%. Moreover, ChOiRe's Steps (i) and (ii) can significantly better fine-tune opinion-aligned models, by up to 18.44%.

1 Introduction

Language models (LMs) are becoming indispensable tools, serving in various roles such as dialogue agents (OpenAI, 2022; Google, 2022), data analysts (Wang et al., 2023a; Cheng et al., 2023), and decision support (Ye et al., 2023). LMs also demonstrate the capability to model distinct opinions which influence response generation on input queries (Bai et al., 2022; Glaese et al., 2022; Santurkar et al., 2023). Unfortunately, the opinions modeled by language models are shaped by the extensive training and feedback data, which are themselves influenced by countless human perspectives, making them inherently challenging to model. As human–AI interactions become common, it becomes imperative to align models with human opinion to meet individual expectations.

Despite the development of alignment frameworks like RLHF (Christiano et al., 2017; Ouyang et al., 2022), aligning large language models (LLMs) with human opinions remains challenging due to the need for significant computational resources and high-quality supervised feedback data, which is difficult to collect. As a result, prompt-based opinion alignment using personas has emerged as a resource-efficient alternative (Perez et al., 2023; Simmons, 2023; Santurkar et al., 2023; Deshpande et al., 2023).

However, even when aligning LLMs with well-represented groups, persona-based prompting methods exhibit low steerability (Santurkar et al., 2023), posing significant concerns and challenges in modeling individual users. Moreover, Hwang et al. (2023) find significant opinion variations among individuals sharing the same demographics, exposing flaws in current group-focused alignment. They argue for individualized models, suggesting to include user's demographic and ideology (which we term as *explicit personae*), and historical opinions (*implicit personae*) for opinion prediction.

While naïvely including explicit and/or implicit personae into the prompt like Santurkar et al. (2023); Hwang et al. (2023) achieves promising results, this personae usage for LLMs is inefficient suffering from multiple limitations, mainly because personae commonly contain noisy and irrelevant information. First, all explicit personae are employed. We contend that only a subset is needed for accurate opinion prediction; including non-relevant personae may act as noise, harming predictive performance. Second, Hwang et al. (2023) utilize the top-K semantically similar opinions with the question (here termed top-K implicit personae). This approach is inefficient, as similar opinions may not offer the most valuable information for prediction. Interestingly, our empirical experiments suggest that LMs may lack sufficient personae evidence with this fixed K — dynamically adjusting K per task can overcome such deficiencies. Finally, while Chain-of-Thought (CoT; Wei et al. 2022; Kojima et al. 2022) enables LMs to perform multi-step reasoning tasks effectively, we surprisingly find that the naïve application of CoT does not help modern LLMs like ChatGPT with opinion alignment

To address the above challenges, we propose ChOiRe[1] (fig. 1), a novel four-step solution for opinion prediction leveraging LLMs' strong data analytic capabilities (Wang et al., 2023a; Cheng et al., 2023). First, an LLM analyzes a target user's explicit personae to discard irrelevant ones. Second, the LLM ranks implicit persona opinions in order of usefulness, selecting the top-K as the most valuable. This surpasses the constraint of using semantic similarity scores. Third, we introduce *Chain-of-Opinion (CoO)*, a designed variant of CoT that allows the LLM to explain and analyze selected explicit personae and top-K implicit personae sequentially. Finally, ChOiRe applies self-consistency over CoO to provision the appropriate amount of user information for opinion inference.

ChOiRe achieves new state-of-the-art (SOTA) in opinion alignment effectiveness and reliability, while using a limited inference budget (appendix C.4). We conduct a thorough analysis to verify our hypotheses concerning explicit and implicit personae and defend our Chain-of-Opinion reasoning methodology. Moreover, ChOiRe's first two steps significantly boost fine-tuning opinion-aligned models. Additionally, ChOiRe generalizes well in missing persona(e) circumstances, and four ChOiRe's steps are also generalizable and motivative for other personalized tasks where the explicit personae and user historical views are available.

2 Related Work

Aligning LMs with Humans. Aligning language models with human behaviour is a recent area of study as alignment can increase user experience satisfaction and utility (Wang et al., 2023c). One line of work develops prompting techniques with user demographic information (e.g., political identity) to encourage LMs to output human-like responses. Argyle et al. (2023) show that by properly conditioning LMs with targeted identity and personality profiles, it is possible to produce biased outputs that strongly correlate with human responses. Furthermore, Simmons (2023) claims that LLMs are moral mimics: by giving models a political identity, they produce texts mirroring the associated moral biases. Despite recent advances, Santurkar et al. (2023) discovered that LMs align poorly with human opinions, as evidenced by model performance on public opinion polls. Hwang et al. (2023) recently propose to incorporate explicit and implicit personae to predict human opinions in new contexts. In §1, we argue that this naïve strategy is suboptimal. ChOiRe overcomes these limitations.

Reasoning with LMs via Prompting. Large-scale model architectures (Devlin et al., 2019; Radford et al., 2019; Brown et al., 2020; Chowdhery et al., 2023; Touvron et al., 2023) have enabled large language models (LLMs) to excel at various NLP tasks using zero- or few-shot prompting (Liu et al., 2023). Notably, Wei et al. (2022); Kojima et al. (2022) propose prominent Chain-of-Thought (CoT) techniques, enabling LLMs to explicate intermediate reasoning steps to solve multi-step reasoning tasks with higher fidelity and efficiency.

Can CoT analyze and predict human opinion effectively? We find that a naïve application of CoT does not help GPT-X models (§5), but that an appropriate modification does. We propose Chain-of-Opinion (CoO) reasoning (§3) that overcomes CoT's limitations in this task. Noting that prompting techniques such as task decomposition (Khot et al., 2023; Zhou et al., 2023) and retrieved-based methods (Yao et al., 2023; Shinn et al., 2023) have been recently introduced, we focus only on the reasoning explanation aspect here given the abstractive and challenging nature of the task.

3 ChOiRe: A Chain of Opinion Framework

Task Formalisation. We follow Santurkar et al. (2023), and formulate the opinion prediction task as multiple-choice question answering. Formally, a benchmark with N data points is denoted as $D = \{\langle T, E, I, q, a\rangle_n\}_{n=1}^{N}$, where T, E and I in-

[1]ChOiRe, **Ch**ain of **O**p**i**nion **Re**asoning, pronounced as the English word "choir".

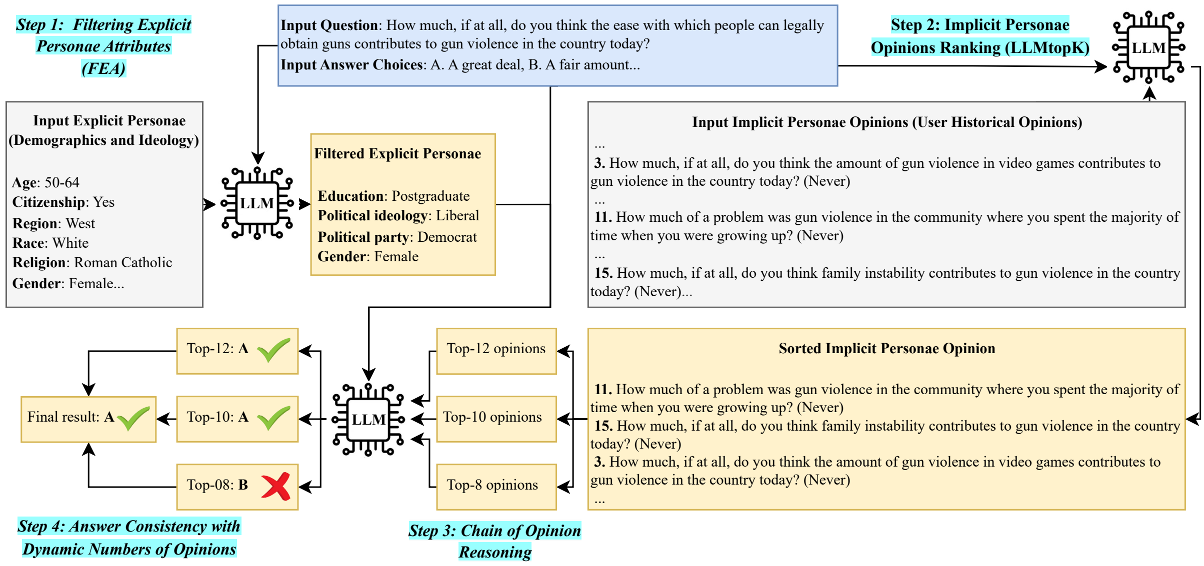

Figure 1: ChOiRe overview, consisting of the four main steps (cyan background), as detailed in §3.

dicate the *(T)opic* of a question q, the *(E)xplicit personae* and *(I)mplicit personae* of the user answering q with opinion a. Following the prior work, E consists of 12 user demographic and ideology metadata attributes, and I contains a number of the user's historical opinions in the format of question–answer pairs. Models then learn to analyze T, E, I, q and predict the opinion a.

Fig. 1 shows an overview of ChOiRe, consisting of four main steps (marked with a cyan background). First, ChOiRe employs an LLM to analyze and select a subset of relevant explicit personae, denoted as $E^{rel} \subseteq E$ for answering the opinion question q. The LLM then assesses the informativeness of the implicit personae (I) in predicting q, selecting the top-K implicit personae (termed *LLMtop-K*). Next, an LLM is prompted to explain the provided explicit E^{rel} and implicit *LLMtop-K* personae sequentially in a *Chain-of-Opinion (CoO)* reasoning strategy. Finally, ChOiRe calls the LLM to predict the opinion a with varying values of K for the top-K implicit personae. ChOiRe chooses the opinion with the highest frequency as the final prediction. We detail the steps below.

3.1 Filtering Explicit Personae Attributes (FEA)

Accounting for explicit personae, which consist of the demographic and ideological metadata attributes of users — such as their `age`, `political view` — are shown to help models characterize and predict human opinions more accurately (Santurkar et al., 2023; Hwang et al., 2023). However, *which personae matter and which do not?* are still open questions. Appendix E.1 shows such an example in full, where, when considering all of explicit personae, the model makes an incorrect prediction while removing unnecessary personae the model made a correct prediction. This may be caused by the LLM's attention mechanism's forcing the model to attend to all input tokens, even irrelevant ones. To filter out unnecessary explicit personae, we ask the LLM to reason and analyze how each persona is helpful for the model to predict the opinion via Chain-of-Thought to output a list of which personae are relevant given the question and the opinion answer choices[2]. Surprisingly, we find that LLMs evaluate more than half of the explicit personae as irrelevant on average. We further conduct human evaluations to verify this finding in §4.

3.2 Implicit Personae Opinions Ranking (LLMtop-K)

LLMs have been sensitive to selected demonstrations and their order in the prompts (Perez et al., 2021; Luo et al., 2023; Gao et al., 2023). For predicting human opinions, we discover that LLMs are also sensitive to the chosen implicit personae opinions as input. Hwang et al. (2023) rank the implicit personae opinions via semantic-similarity scores and select top-K. This strategy is suboptimal because the top-ranked opinions in terms of semantic similarity may not be the ones that provide the most supportive information for the models to predict opinions (appendix E.2). As LLMs are shown to be good data analysts (Wang et al., 2023a; Cheng

[2]We provide our FEA prompt in Appendix C.1

et al., 2023), we propose to address the above challenge by utilizing LLMs to analyze and rank the implicit personae opinions in usefulness descending order. Our finding is that despite the output rankings from LLMs varying with different input orders of implicit personae opinions, the sets of LLMtop-K opinions overlap by a good coefficient when K is large enough (≥ 8) (fig. 5). Therefore, we propose to input the implicit personae opinions to LLMs in a random order to make our method more versatile. We also examine the case where we input the opinions in the semantic similarity order. We illustrate the prompt template in appendix C.2. By performing this step, our proposed method supports the usefulness of opinions in predicting the test opinions, rather than the semantic similarity. We term this method as *LLMtop-K*.

3.3 Chain-of-Opinion Reasoning (CoO)

Wei et al. (2022); Kojima et al. (2022) introduce few-shot and zero-shot Chain-of-Thought (CoT) prompting strategies demonstrating that by reasoning step-by-step, LLMs can achieve promising results on complex tasks. However, the sampled reasoning steps can be inconsistent, leading to possibly different outcomes (Wang et al., 2023b). Furthermore, it is little known how the models perceive multiple implicit personae opinions, especially when many opinions are input, *which one(s) the models used, which one(s) they didn't for predicting the opinion?* Our preliminary experiments with CoT (§6.1 and appendices E.3 and E.4) reveal that the CoT explanations can vary frequently based on different subsets of opinions mentioned in their explanations, leading to diverse final answers, especially when the decoding temperature is relatively high (≥ 0.6). To mitigate this issue, we propose to instruct the LLMs to analyze the given explicit and implicit personae one by one before concluding the prediction via simply adding `"explaining and analyzing how each of the Opinions and Demographic Information supports the question"` into the prompt instruction. Given an LLM that can follow human instructions well such as ChatGPT (OpenAI, 2022), this addition offers two notable advantages. First, for each question, we ensure that the model explains and analyzes the provided personae one by one without missing any, possibly resulting in more thorough predictions. Second, this method helps the model to output more consistent reasoning explanations, enhancing its reliability (§6.1).

3.4 Answer Consistency with Dynamic Numbers of Opinions

Prior work (Hwang et al., 2023) fixes the number of implicit personae opinions for prediction to $K = 8$. However, this approach occasionally results in models generating `"...the answer cannot be determined."` (table 4 and appendix E.5). We attribute this to insufficient user implicit personae opinions provided. Inspired by *Self-Consistency (SC)* (Wang et al., 2023b), our approach involves sampling multiple answers using different K values for a given question. The most frequent answer, along with the explanation of the first correct answer, becomes the final prediction. Our method is distinct from SC since SC samples multiple answers with a fixed prompt. We experiment with $K \in \{8, 10, 12\}$ for efficiency (appendix C.4).

4 Evaluation

Dataset. We experiment on OpinionQA dataset (Santurkar et al., 2023) — the only opinion QA dataset to date consisting of both user explicit and implicit personae designed for the assessment of alignment between LMMs' opinions and human participants, encompassing a diverse range of 60 US demographic groups. It covers 60 US demographic groups, with 15 topics, each comprising around 100 questions, gathered from $5,340$ users.

Dataset Preprocessing. Due to limited resources, we randomly sample 25 users per topic for our experiments. For each user, we follow Hwang et al. (2023) to use 20% of the implicit questions as the implicit persona. For the remaining 80% implicit questions, we randomly select a maximum of 15 implicit questions for testing. Our sampling method results in a total of 375 users and $5,603$ implicit evaluation question–answer pairs. Our subset is highly representative because we gather 25 users from every topic and 15 questions per user. Rigorous statistical tests further validate the significance of our results which align closely with Hwang et al. (2023) testing on a larger subset using InstructGPT.

Prompting Baselines. We use both closed-source ChatGPT (OpenAI, 2022), ChatGPT-Instruct (OpenAI, 2023a), GPT-4 (OpenAI, 2023b), and open-source Mistral-7B-Instruct-v.02 (Jiang et al., 2023) as our LLMs, and compare ChOiRe with 5 prompting methods: *(1) W/o persona*, where LLMs are evaluated without user historical opinions, ideology, or demographic data; *(2) Demographic + Ideology + top8 Opinions* (termed *DIO-top8*), introduced by Hwang et al. (2023) demonstrating that integrating explicit and implicit personae enhances user opinion modeling and prediction, achieving state-of-the-art results on OpinionQA at that time; *(3) DIO-top8 + CoT* is the Chain-of-Thought (CoT) prompting (Kojima et al., 2022) version of *DIO-top8* involving appending `"answer the following question step-by-step"` to prompts, aiming to explore whether CoT improves model performance in this task; *(4) DIO-top8 + SC* is the baseline which we apply the Self-Consistency technique with CoT (Wang et al., 2023b) to *DIO-top8* to select the most frequent answer generated by the model as the final opinion prediction; *(5) DIO-top8 + Self-refine* (Madaan et al., 2023) interactively feedbacks and refines the answers by LLMs. We do not experiment with InstructGPT (Ouyang et al., 2022) like Hwang et al. (2023) since this model is going to be deprecated and replaced by ChatGPT-Instruct. For GPT-4, we only run the main experiment and we use ChatGPT for FEA and LLMtop-K steps due to our limited budget. All the prompts and costs are in appendix C, implementation details in appendix A.1, and more baselines in appendix B.2.

Fine-tuning Baselines. We further investigate whether ChOiRe's FEA and LLMtop-K steps (§3) also improve fine-tuning for opinion-aligned models. We first create the fine-tuning data by using ChatGPT to perform ChOiRe's FEA and LLMtop-K steps ($K = 8$) on a training set of $30,000$ samples randomly selected from OpinionQA which are different from our $5,603$ test ones. We then fine-tune and evaluate GPT-2 models (base, large) (Radford et al., 2019) and FlanT5 models (base, large) (Chung et al., 2022). Fine-tuning details are provided in appendix A.1.

Metrics. We employ *Accuracy* and *Collapsed Accuracy*[3] as the automatic evaluation metrics following Hwang et al. (2023). It is worth noting that *Precision/Recall/F1* is not applicable in our task, since the numbers of answer choices are not the same for all the OpinionQA samples. In addition, human evaluations are crucial due to the absence of automated metrics assessing LLMs' performance

[3]is a relaxed accuracy wherein the choices of MCQ questions ($\geq$ 4 choices) are collapsed to become 2 choices.

Model	ChatGPT	ChatGPT-Inst	GPT-4	Mistral-7B-Ins.-v0.2
W/o persona	46.60/65.72	44.91/63.60	-	41.24/59.54
DIO-top8	50.22/69.21	51.95/71.16	57.98/76.86	44.16/62.47
DIO-top8 + Self-refine	43.14/65.33	42.71/62.98	-	36.23/55.06
DIO-top8 + CoT	49.96/69.05	51.90/71.51	-	52.25/71.95
DIO-top8 + SC	50.58/69.66	52.06/71.87	-	53.14/72.88
DIO-top8 + FEA	50.64/69.85	52.63/72.30	-	44.99/64.09
DIO-top8 + CoO	50.97/70.22	52.08/71.65	-	53.79/73.59
DIO-LLMtop8	51.03/70.31	52.80/72.60	-	45.86/64.98
DIO-LLMtop8 + FEA	51.19/70.69	52.97/72.84	-	45.23/64.73
DIO-LLMtop8 + FEA + CoO	51.90/71.57	53.01/72.91	59.02/78.70	54.21/74.09
ChOiRe	**52.21†/72.09†**	**53.26†/73.26†**	**59.30†/78.82†**	**54.43†/74.34†**
% Improvements	+3.22/+3.49	+2.52/+1.93	+2.28/+2.55	+2.42/+2.00

Table 1: Accuracy/Collapsed Accuracy on ChatGPT, ChatGPT-Instruct, and GPT-4. *FEA* is our first step, *Filtering Explicit Attributes*. *LLMtop8* is the second step, *CoO* is *Chain-of-Opinion reasoning*. Improvements are calculated with the best baseline. † denotes our model outperforms baselines significantly with p-value < 0.01 under t-test (table 8).

Model	GPT-2-base	GPT-2-large	FlanT5-base	FlanT5-large
W/o persona	**41.14/58.87**	21.94/39.11	48.98/68.33	39.83/58.43
DIO-top8	21.23/38.64	24.94/42.22	55.00/74.98	54.94/74.79
DIO-top8 + FEA	22.62/40.97	25.65/45.21	55.78/75.34	58.77/77.26
DIO-LLMtop8	22.65/41.12	28.86/47.60	**57.97†/77.46†**	58.20/77.56
DIO-LLMtop8 + FEA	25.05/44.41	**29.54†/48.66†**	57.45/77.13	**59.00†/78.46†**
% Imp. over DIO-top8	+17.99/+14.93	+18.44/+15.25	+5.40/+3.30	+7.38/+4.90

Table 2: Performance of fine-tuned baselines with our proposed *FEA* and *LLMtop8* steps preprocessed by ChatGPT. † denotes our model significantly outperforms baselines with p-value < 0.01 under t-test (table 8).

in FEA, LLMtop-K and CoO steps of ChOiRe. Therefore, we conduct our human assessments to address these research questions: *(1) LLMs' effectiveness in filtering unnecessary explicit personae*; *(2) LLMs' proficiency in ranking implicit personae opinions*; *(3) LLMs' ability to explain answers via CoO*. To this end, we randomly select 100 answers generated by ChOiRe with ChatGPT, ChatGPT-Instruct, GPT-4, Mistral. We then hire 3 excellent undergraduates who are native English speakers as annotators. For FEA and LLMtop-K steps, each annotator is instructed to rate on a 1-3 scale (3 is best) via the **Satisfaction** criterion defined as how well the algorithm of LLMs performs in filtering/ranking, subjectively. To answer (3), we use two criteria named **Reasonableness** measuring how well the LLMs reason with the CoO explanations, and **Follow the Instruction** assessing the capability of LLMs in following our instruction to explain and predict the opinions. Three annotators are also guided to rate the criteria on a 1-3 scale. Each metric's final score is the average of three annotators' scores. The scoring instructions are in appendix D.1 and the inter-annotators' agreement is assessed by Kripp's alpha (Krippendorff, 2011).

5 Main Results

Overall Prompting Results. Table 1 shows our macro experimental outcomes. We derive 4 main observations in this task. First, ChOiRe improves the best among baselines significantly with

3.22%, 2.52%, 2.28%, 2.42% accuracy for ChatGPT, ChatGPT-Instruct, GPT-4, and Mistral. It establishes a strong SOTA result with GPT-4, surpassing previous SOTA DIO-top8 with InstructGPT achieving 53.74% (Hwang et al., 2023) by a notable margin. Notably, in the case of GPT-4, we utilize ChatGPT for FEA and LLMtop-K steps, showcasing the strength of a weaker model that enhances a stronger one. Second, we see that Accuracy and Collapsed Accuracy have the same trend, and ChOiRe also achieves the SOTA on Collapsed Accuracy with the highest improvement of 3.94% observed with ChatGPT. Third, naïve CoT (`"answer the following question step-by-step"`) helps Mistral but slightly harms ChatGPT and ChatGPT-Instruct with *DIO-top8* (Hwang et al., 2023). On the other hand, SC improves all models. Therefore, we attribute CoT's limitation to the inconsistency of its explanations (3). Meanwhile, ChOiRe with CoO consistently attains improvements, verifying the effectiveness of explicitly requiring the model to analyze all the personae. Finally, ChatGPT, ChatGPT-Instruct, and Mistral show improvements by selecting only $4.79/12$ and $5.59/12$, $8.83/12$ explicit personae on average, respectively. This suggests that over half of explicit personae may be noisy for models to predict opinions.

Fine-grained Prompting Results. Diving deeper into the benchmark topics in table 5, ChOiRe achieves SOTA results in $8/15$, $8/15$, $11/15$, $13/15$ topics for ChatGPT and ChatGPT-Instruct, GPT-4, and Mistral. The improvements are especially huge for some topics. For example, compared with the best among baselines, it improves GPT-4 up to 12.08% accuracy on *Views on gender*, ChatGPT up to 9.82% on *Economic Inequality*. We also specifically compare ChOiRe with the best baseline DIO-top8 + SC in appendix B.6, showing $8/12$ improvements for ChatGPT and ChatGPT-Instruct. We further plot the accuracy distribution over users of ChOiRe, specifically for ChatGPT in fig. 4. We see that the majority accuracy is 0.5, with a few users scoring zero and over 20 achieving perfection.

Fine-tuning Results. Table 2 presents our fine-tuning outcomes. Notably, leveraging the ChOiRe's FEA and LLMtop-K steps on the fine-tuning data yields substantial enhancements for GPT-2-large and FlanT5-large, showcasing relative accuracy improvements of 18.44% and 7.38% respectively. Remarkably, ChOiRe's FEA and LLMtop-K steps bring FlanT5-large's performance on par with GPT-4, despite GPT-4's significantly stronger capability. Furthermore, ChOiRe's LLMtop-K proves particularly beneficial for enhancing FlanT5-base. Surprisingly, GPT-2-base performs well even without user demographic and ideological information, possibly due to potential contamination (Sainz et al., 2023) with public polling data from OpinionQA.

Model	FEA Satis.	LLMtopK Satis.	Rea.	Foll. Inst.
ChatGPT	2.56 (Kα' 0.74)	**2.32** (Kα' 0.68)	2.90 (Kα' 0.88)	**2.95** (Kα' 0.90)
ChatGPT-Inst.	**2.64** (Kα' 0.71)	2.28 (Kα' 0.65)	2.92 (Kα' 0.90)	**2.95** (Kα' 0.87)
GPT-4	-	-	**2.95** (Kα' 0.91)	2.21 (Kα' 0.77)
Mistral-7B-Ins.-v0.2	2.31 (Kα' 0.65)	2.12 (Kα' 0.64)	2.66 (Kα' 0.68)	2.16 (Kα' 0.55)

Table 3: Human evaluation results. Kα' is Kripp's alpha.

Human Evaluation Results. Our human evaluation results in table 3 reveal three key findings. First, ChatGPT and ChatGPT-Instruct achieve similar performance in filtering explicit personae and ranking opinions, while Mistral achieves lower results. While ChatGPT excels slightly in ranking, ChatGPT-Instruct performs slightly better in explicit personae selection. Three models proficiently filter unnecessary explicit personae, but ranking opinions poses a more challenging task intuitively and empirically, with a common error being the inconsistent relevance ranking of opinions, sometimes misplacing high-level relevance. Second, four models effectively generate reasonable thoughts leading to the final answer, and GPT-4 performs the best. Finally, ChatGPT and ChatGPT-Instruct follow our instructions to explain and analyze the explicit and implicit personae provided one by one with CoO significantly better than GPT-4 and Mistral, achieving nearly perfect scores of 3. We hypothesize that this is because ChatGPT and ChatGPT-Instruct excel in following instructions, while GPT-4 is optimized for completing texts.

6 Discussion

We discuss the main analyses in this section. Extra important analyses are presented in appendix B.

6.1 Methodology Analysis

Ablation of FEA. To gauge the impact of filtering unnecessary explicit personae (FEA) on performance, we experiment with applying FEA exclusively to the baseline DIO-top8 (Hwang et al., 2023), denoted as *DIO-top8 + FEA* in table 1. The results indicate enhancements with DIO-top8

+ FEA achieving a 0.8%, 1.3%, 1.9% accuracy performance boost on ChatGPT, ChatGPT-Instruct, and Mistral respectively. This underscores the effectiveness of eliminating irrelevant explicit personae in improving the models' ability to understand and predict human opinions.

FEA via Topics. To understand the explicit personae filtered by LLMs across various topics, we document the top-3 removed personae in appendix B.8. We observe that `"Citizenship"` is consistently the most frequently removed attribute, followed by `"Race"`. This could be due to LLMs treating these as sensitive information, prioritizing respect and unbiased text generation. Another explanation may be the lack of correlation between citizenship/race and opinions in the US-centric OpinionQA dataset. Additionally, we also see that ChatGPT often categorizes `“Marital status”` as non-useful, ChatGPT-Instruct commonly removes `“Frequency of religious attendance”`, and `“Gender”` got removed by Mistral, revealing potential biases in LLMs.

LLMtop-K versus Top-K. From table 1, DIO-LLMtop8 outperforms DIO-top8 by 1.6%, 1.6%, 3.8% accuracy on ChatGPT, ChatGPT-Instruct, Mistral confirming that prioritizing meaning and usefulness improves opinion prediction. One possible explanation for this can be the orders ranked by semantic similarity scores only consider ranking with the input questions (Hwang et al., 2023), while our orders consider both input questions and their answer choices (fig. 1). We further explore two key aspects: *(1) The agreement of LLM-orders and semantic similarity orders*, and *(2) Points of maximum disagreement between these orders*. To measure the ranking agreements, we calculate Kendall's Tau correlation coefficient (Kendall, 1938) between the orders generated by ChatGPT, ChatGPT-Instruct, and Mistrial and orders sorted by semantic similarity scores, and the results are presented in fig. 6 and fig. 7. Surprisingly, for ChatGPT and ChatGPT-Instruct, we find that the two ranking orders have minimal monotonous relations with means approximating 0 and low standard deviations showing no agreement. For Mistral, we find a low agreement with a mean of 0.43 score. These low and no agreements further verify that ranking by usefulness can be very different from ranking by semantic similarity. We also deep dive into cases with notable order variations to address (2). Appendix E.2 illustrates one such case in the `"Guns"` topic. We derive three observations. First, not all top-8 opinions by semantic similarity scores help predict the opinion. For example, the 16-th opinion, despite having a relatively high semantic similarity score with the question which might offer some perspective on the prevalence of guns in the user's community during the upbringing, is less directly related to the question. This is similar to the 18-th opinion which is also less relevant. Meanwhile, several important opinions are deselected by the semantic-similarity-based method, such as the $6, 3, 4, 10$-th ones, which are chosen by the LLM. The 6-th one is critical, and directly relevant because it assesses the person's attitude toward safety measures related to gun ownership. Finally, by using LLMtop-K order, the model predicts the opinion accurately, whereas the semantic similarity order leads to an incorrect prediction.

Opinions Order Analysis of LLMtop-K Step. The performance difference of DIO-top8 and DIO-LLMtop8 in table 1 highlights that LLMs are sensitive to the chosen implicit personae opinions. An important question arises: *Are LLMs also affected by the input order of implicit persona opinions in the ranking step (§3)*? Our discovery confirms sensitivity, but with reasonable overlap when K is sufficiently large. We randomly select 300 questions, shuffle implicit persona opinions four times with different seeds, and record four LLM ranking outputs for each. We also collect one more LLM ranking output by feeding implicit personae opinions in semantic similarity order. For each $K \in \{1, 2, ..., 20\}$, we calculate the pairwise Overlap coefficient (Vijaymeena and Kavitha, 2016) among the five ranking outputs, averaging them as the **LLM ranking consistency score** for each K. The scores, shown in fig. 5, indicate that for $K \geq 8$, the ranking outputs overlap well with a score of $\geq .6$ for both models. Despite this, *is there substantial variance in model performance across random seeds*? Our findings reveal no significant variance, with the variants statistically outperforming the baseline DIO-top8. Specifically, we assess ChatGPT and Mistral with DIO-LLMtop8 on 3 out of 4 random seeds, detailed in appendix B.9. The results demonstrate relatively small standard deviations in their performance, and critical values of 99% CI of DIO-LLMtop8 under t-test for both models surpass DIO-top8, confirming that LLM-top8's effectiveness is not due to randomness.

Model	ChatGPT	ChatGPT-Inst	GPT-4	Mistral
% of ITA of DIO-LLMtop8 + FEA + CoO	0.61	1.32	9.71	0.00
DIO-LLMtop8 + FEA + CoO	51.90	53.01	59.02	54.21
% of ITA of DIO-LLMtop10 + FEA + CoO	0.12	1.01	5.44	0.00
DIO-LLMtop10 + FEA + CoO	51.55	52.74	58.88	53.88
% of ITA of DIO-LLMtop12 + FEA + CoO	0.00	0.66	3.12	0.00
DIO-LLMtop12 + FEA + CoO	51.60	52.31	59.11	52.96
ChOiRe	**52.21**	**53.26**	**59.30**	**54.43**

Table 4: Extra analysis on ChatGPT, ChatGPT-Instruct, GPT-4, and Mistral. ITA stands for "Impossible To Answer".

CoO versus CoT. Table 1 indicates that Chain-of-Thought (CoT) (Kojima et al., 2022) slightly harms baseline DIO-top8 performance for ChatGPT and ChatGPT-Instruct. Conversely, our Chain-of-Opinion reasoning (CoO) enhances overall performance for all models. To investigate the consistency of CoT and CoO, we design an experiment with ChatGPT, DIO-top8 where we randomly select 100 question-answer pairs and sample 5 answers per pair using CoT and CoO, at 3 different temperatures 0.3, 0.6, 0.9. For each prompting technique, we measure the percentage of questions that all 5 answers sampled have the same result, as the consistency score. The results are illustrated in appendix B.5-fig. 2 showing that CoO brings better consistent answers compared to CoT, especially when the temperature is high verifying CoO potentially enhances the reliability of LLMs.

Dynamic Numbers of Opinions Analysis. Table 4 illustrates our analysis answering two research questions: *(1) How frequently can't LLMs answer the question?* and *(2) How do LLMs perform when more opinions than $K = 8$ are provided?* Our findings show that, firstly, with 8 opinions, GPT-4 exhibits the highest percentage of unanswered questions, while Mistral answers all the questions. Secondly, increasing the number of opinions beyond 8 reduces this percentage across models, confirming our hypothesis regarding the lack of implicit personae opinions when fixing $K = 8$ in §3. Lastly, while including more opinions could harm the performance of models, our answer consistency strategy enables LLMs to achieve the best results across K values.

6.2 Error Analysis

FEA Misses Key Explicit Personae. Despite showing promising results in removing unuseful explicit personae depicted in table 3, we observe that LLMs sometimes misselect relevant personae. One such example is the top-left of appendix E.6. We observe that in this case, our annotators can't grade a high FEA satisfaction score because "`Education`" and "`Age`" are also two important personae as they can influence one's understanding of workplace dynamics significantly, which are deselected by ChatGPT.

LLMtop-K Opinions Include Less Relevant Ones. While LLMs generally demonstrate a commendable ability to rank implicit opinions by usefulness, as exemplified in appendix E.2, we also observe they frequently include less relevant, or even irrelevant opinions to the ranked list such as in appendix E.6-bottom. We attribute this to the challenge of this task, even for humans it might require substantial cognitive effort.

LLMs May Not Follow the Instructions. Although ChatGPT and ChatGPT-Instruct demonstrate a robust ability to adhere to our instructions for opinion prediction via CoO, the same level of proficiency is not observed in Mistral and GPT-4, as shown in appendix E.6-top-right. We posit this disparity arises from the fact that ChatGPT and ChatGPT-Instruct excel in comprehending and executing human instructions, while GPT-4 excels primarily in generating coherent text.

6.3 ChOiRe's Generatization

We discuss two fundamental questions regarding the generalization of ChOiRe: *(1) How does ChOiRe perform in the situations of missing persona(e)?* and *(2) Are ChOiRe's steps generalized to other tasks?* Our full discussions are provided in appendix B.1. In summary, we find that ChOiRe's steps generalize to other tasks, and missing persona(s) situations better than the baselines.

7 Conclusion

We propose ChOiRe, a novel four-step solution framework addressing the problem of effectively employing personae with LLMs for opinion prediction. We further introduce Chain-of-opinion reasoning and answer consistency over variable numbers of input implicit personae guiding the models to derive thorough predictions. ChOiRe achieves strong SOTA results with limited inference calls, demonstrating its strong effectiveness. Additionally, Steps (i) and (ii) of ChOiRe significantly improve the fine-tuning of opinion-aligned models. We strongly suggest that our method should only be used for positive moral intents, avoiding making LLMs echo chambers (Vicario et al., 2016). In the future, we will focus on developing frameworks that utilize personae more efficiently.

Limitations

One limitation of our proposed ChOiRe framework is that it requires the LLMs to have a good capability in following human instructions to solve tasks such as selecting explicit personae, ranking historical opinions, and explaining personae and opinions one by one via CoO. However, we foresee that this limitation is going to be overcome by cutting-edge AI language models, in the present and near future. Additionally, our method also utilizes user's personal information from explicit and implicit personae, which may be sensitive to some audiences and not be fully available in the real world. However, to what extent is the personal information provided, our ChOiRe is still able to offer reasonable opinion predictions since it is not constrained by the number of provided explicit personae, or the number of user historical opinions (see appendix B.1).

Ethical Considerations

Characterizing and predicting human opinions with LLMs can be directly applied to personalize and align machines to users' values, and cultural beliefs. Nonetheless, there exist unwanted situations when LLMs with our techniques can be misused for unethical purposes and biased opinions.

Bias Amplification and Fairness. A personalized LLM allows users to reinforce their existing beliefs and potentially amplify biased or unethical perspectives, leading to the creation of echo chambers (Vicario et al., 2016). This can ultimately harm users by reinforcing polarized or undesirable views. To mitigate this issue, the Chain-of-Opinion (CoO) reasoning from our proposed ChOiRe involves presenting user demography or ideology group responses alongside personalized answers. Additionally, CoO can encourage users to reflect on their previous viewpoints.

Privacy and Consent. Users may not always be aware of or have control over the extent of personalization applied to the content they receive. Therefore, empowering users to have control over AI-generated opinions is essential. Users should be able to customize and adjust the explicit and implicit personae used for opinion prediction. This customization can help mitigate potential biases and provide individuals with AI-generated opinions that align more closely with their values and preferences.

Human Evaluation. Through human evaluations, we observe that our proposed method does not generate any discriminatory, insulting responses. We validate the intermediate steps of our proposed ChOiRe by human evaluation which involves manual labor. We hire annotators to score, and the hourly pay is set to $15, which is higher than the local statutory minimum wage. Therefore, we do not anticipate any major ethical concerns raising from human evaluations.

References

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

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional ai: Harmlessness from ai feedback.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Liying Cheng, Xingxuan Li, and Lidong Bing. 2023. Is GPT-4 a good data analyst? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9496–9514, Singapore. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2023. Palm: Scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24:240:1–240:113.

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

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Dorottya Demszky, Nikhil Garg, Rob Voigt, James Zou, Jesse Shapiro, Matthew Gentzkow, and Dan Jurafsky. 2019. Analyzing polarization in social media: Method and application to tweets on 21 mass shootings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2970–3005, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Xuan Long Do, Bowei Zou, Liangming Pan, Nancy F. Chen, Shafiq Joty, and Ai Ti Aw. 2022. CoHS-CQG: Context and history selection for conversational question generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 580–591, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Shuzheng Gao, Xin-Cheng Wen, Cuiyun Gao, Wenxuan Wang, and Michael R. Lyu. 2023. Constructing effective in-context demonstration for code intelligence tasks: An empirical study. *CoRR*, abs/2304.07575.

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*.

Google. 2022. Bard: A conversational ai tool by google.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

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

EunJeong Hwang, Bodhisattwa Prasad Majumder, and Niket Tandon. 2023. Aligning language models to user opinions. *CoRR*, abs/2305.14929.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

M. G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2023. Decomposed prompting: A modular approach for solving complex tasks. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Sunghwan Mac Kim, Qiongkai Xu, Lizhen Qu, Stephen Wan, and Cécile Paris. 2017. Demographic inference on Twitter using recursive neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 471–477, Vancouver, Canada. Association for Computational Linguistics.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9):195:1–195:35.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Man Luo, Xin Xu, Zhuyun Dai, Panupong Pasupat, Seyed Mehran Kazemi, Chitta Baral, Vaiva Imbrasaite, and Vincent Y. Zhao. 2023. Dr.icl: Demonstration-retrieved in-context learning. *CoRR*, abs/2305.14128.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*.

OpenAI. 2022. Introducing chatgpt.

OpenAI. 2023a. Gpt-4 api general availability and deprecation of older models in the completions api.

OpenAI. 2023b. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True few-shot learning with language models. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 11054–11070.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemi Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. 2023. Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13387–13434, Toronto, Canada. Association for Computational Linguistics.

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

Delip Rao, Michael Paul, Clay Fink, David Yarowsky, Timothy Oates, and Glen Coppersmith. 2011. Hierarchical bayesian models for latent attribute detection in social media. In *Proceedings of the international AAAI conference on web and social media*, volume 5, pages 598–601.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore. Association for Computational Linguistics.

Shigeyuki Sakaki, Yasuhide Miura, Xiaojun Ma, Keigo Hattori, and Tomoko Ohkuma. 2014. Twitter user gender inference using combined analysis of text and image processing. In *Proceedings of the Third Workshop on Vision and Language*, pages 54–61, Dublin, Ireland. Dublin City University and the Association for Computational Linguistics.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004. PMLR.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik R Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Gabriel Simmons. 2023. Moral mimicry: Large language models produce moral rationalizations tailored to political identity. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 282–297, Toronto, Canada. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Michela Del Vicario, Gianna Vivaldo, Alessandro Bessi, Fabiana Zollo, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. 2016. Echo chambers: Emotional contagion and group polarization on facebook. *CoRR*, abs/1607.01032.

MK Vijaymeena and K Kavitha. 2016. A survey on similarity measures in text mining. *Machine Learning and Applications: An International Journal*, 3(2):19–28.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023a. Is ChatGPT a good NLG evaluator? a preliminary study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11, Singapore. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

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

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Lanling Xu, Junjie Zhang, Bingqian Li, Jinpeng Wang, Mingchen Cai, Wayne Xin Zhao, and Ji-Rong Wen. 2024. Prompting large language models for recommender systems: A comprehensive framework and empirical analysis. *arXiv preprint arXiv:2401.04997*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Yining Ye, Xin Cong, Yujia Qin, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2023. Large language model as autonomous decision maker. *CoRR*, abs/2308.12519.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V. Le, and Ed H. Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

A Baselines

A.1 Baselines Implementation Details

Prompting. ChatGPT *(gpt-3.5-turbo-0613)*, ChatGPT-Instruct *(gpt-3.5-turbo-instruct-0914)*, GPT-4 *(gpt-4-0613)* are called via OpenAI API with chat, text, text completion mode respectively at a temperature of 0.3. Mistral-7B-Instruct-v0.2 is called via HuggingFace interface[4]. We use Nucleus Sampling (Holtzman et al., 2020) with a $p = .95$ as our decoding strategy. To obtain the embeddings of opinions for semantic similarity scores' computations, we use OpenAI's *text-embedding-ada-002* model with its default setting, following Hwang et al. (2023). For each sample, ChOiRe requires 5 inference calls, 2 for FEA and LLMtop-K steps, and 3 for $K \in \{8, 10, 12\}$. Therefore, to have a fair comparison with our method, we sample 5 answers for the Self-Consistency baseline, and 2 rounds of feedback-edit for Self-refine baseline, for each question.

Fine-tuning. We fine-tune GPT-2 (Radford et al., 2019) and FlanT5 (Chung et al., 2022) base and large sizes to verify that ChOiRe's FEA and LLMtop-K steps (§3) also help to build better opinion-aligned models. Both models with two different sizes are initialized from public pre-trained checkpoints on the Transformers library (Wolf et al., 2020) of HuggingFace. We use a learning rate of $1e-5$ for FlanT5, and $5e-5$ for GPT-2, and AdamW (Loshchilov and Hutter, 2018) as our optimizer with a warm-up of 100 steps. FlanT5 variants are trained on $50K$ iterations, and evaluations and checkpoint-savings are done for each 1000 steps. GPT-2 base model is trained on 15 epochs and evaluated every 300 steps, while GPT-2 large is trained on only 5 epochs, and the checkpoints are evaluated every 300 steps. All the models are fine-tuned on a single A100 80GB GPU. We use a window size of 1024 for both models, and Nucleus Sampling (Holtzman et al., 2020) with a $p = .95$ as our decoding strategy, same as API/inference models. The input format for both models is "Input: explicit_persona <SEP> implicit_persona <SEP> question <SEP> answer_choices; Output: correct_answer" for with persona cases, and "Input: question <SEP> answer_choices; Output: correct_answer" for without persona case. The "correct_answer" is an actual text correct answer like "Yes/No", unlike API/inference models where we use "A/B/C/D". We find that fine-tuning with the textual correct answer yields significantly better results compared to "A/B/C/D", while prompting with "A/B/C/D" for API/inference models achieve slightly better results compared to textual output.

[4] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

B Extra Analysis

B.1 ChOiRe's Generalization

How Does ChOiRe Perform If Not Enough Personae Given? We conduct experiments with ChOiRe in Table 6 using ChatGPT (gpt-3.5-turbo-1106) under the following conditions of not having enough explicit and/or implicit personae. Under these cases, ChOiRe is simplified. Specifically:

- **ChOiRe w/o any personae (Step 4 is used)**: Without any personae given, Steps 1 (FEA), 2 (LLMtop-K), 3 (CoO) of ChOiRe are deactivated, while Step 4 is still in use. In this case, Step 4 is simplified to be Self-consistency (Wang et al., 2023b). We observe a 2.55% improvement over the baseline *DIO-top8 w/o persona*.

- **ChOiRe w/o explicit persona (Steps 2, 3, 4 are used)**: Without explicit personae, Step 1 FEA of ChOiRe is deactivated, while Steps 2, 3, 4 are in use. Compared with the baseline *DIO-top8 w/o explicit persona*, it is observed a significant gain up to 4.7% using ChatGPT.

- **ChOiRe w/o implicit persona (Steps 1, 4 are used)**: Without implicit persona, Steps 2 (LLMtop-K) and 3 (CoO) of ChOiRe are deactivated, while Steps 1 (FEA) and 4 (Self-consistency) are utilized. From Table 6, ChOiRe significantly improves *DIO-top8* baseline by a 5.95% Accuracy.

In summary, under the missing persona(e) circumstances, we observe that ChOiRe can generalize well and better than the previous SOTA method *DIO-top8* of Hwang et al. (2023).

ChOiRe's Generalization to Other Tasks. Each of ChOiRe's steps holds a similar philosophy with multiple prior studies proven effective in other personalized tasks. We specify them below:

- **ChOiRe's Step 1: FEA.** Filtering irrelevant user profile attributes for better classification

Model	ChatGPT/ChatGPT-Inst/GPT-4/Mistral				
	Guns	**Auto. & driverless vehicles**	**Views on gender**	**Com. types & sex. harassment**	**Race**
W/o persona	53.07/37.30/——-/30.48	47.73/48.26/——-/41.72	50.53/42.94/——-/37.39	**47.73**/41.67/——-/29.34	41.95/45.28/——-/37.55
DIO-top8	53.87/57.00/60.39/44.73	45.33/44.78/**53.22**/41.72	53.21/52.15/63.73/40.09	43.47/**45.24**/42.86/35.45	43.06/44.65/**55.17**/41.11
DIO-top8 + CoT	54.55/52.33/——-/55.48	47.22/46.77/——-/49.00	48.11/**57.67**/——-/54.28	42.39/42.26/——-/42.01	**45.63**/43.40/——-/49.78
DIO-top8 + SC	54.40/52.85/——-/56.57	43.73/48.26/——-/52.31	55.61/56.44/——-/56.30	45.33/40.48/——-/42.01	45.00/43.40/——-/50.00
ChOiRe	**57.06/58.21/63.37/58.00**	**49.25/51.92**/50.00/**53.75**	**59.23**/53.07/**71.43/57.78**	39.88/44.14/**47.96/42.08**	42.77/**47.28**/50.57/**51.44**
	Gender & Leadership	**America in 2050**	**Trust in science**	**Biomedical & food issues**	**Misinformation**
W/o persona	**53.13**/50.83/——-/43.51	39.73/39.13/——-/41.95	50.40/47.29/——-/48.34	53.87/53.63/——-/53.21	46.93/40.38/——-/53.63
DIO-top8	48.27/54.70/**65.55**/50.23	46.93/46.20/43.70/35.14	54.93/61.58/61.54/51.65	52.27/55.86/58.03/52.78	49.33/52.11/52.71/50.77
DIO-top8 + CoT	48.58/50.83/——-/55.79	43.05/48.91/——-/43.76	54.10/65.02/——-/58.28	**56.91/57.54**/——-/57.08	**49.57/53.99**/——-/53.19
DIO-top8 + SC	49.07/53.60/——-/**57.87**	45.87/47.83/——-/46.03	56.27/**65.52**/——-/58.94	53.07/**57.54**/——-/**58.58**	45.00/53.52/——-/**53.85**
ChOiRe	52.22/**57.78**/63.03/**57.87**	**49.46/48.99/45.37/47.50**	**56.43**/55.50/**68.46/60.37**	54.75/57.26/**61.61/58.58**	46.45/53.62/**57.36/53.85**
	Privacy & Surveilance	**Family & Relationships**	**Economic inequality**	**Global attitudes**	**Political views**
W/o persona	43.24/40.28/——-/33.64	47.06/44.36/——-/46.08	43.67/49.15/——-/34.07	46.13/46.71/——-/40.42	40.80/48.95/——-/46.20
DIO-top8	53.24/47.22/47.73/43.31	57.22/57.89/62.50/47.42	45.60/**51.98**/63.81/41.87	**49.60**/57.23/**66.67**/41.27	**56.80**/46.85/62.07/44.13
DIO-top8 + CoT	53.38/47.22/——-/56.91	59.57/55.64/——-/54.36	47.65/**51.98**/——-/51.45	46.42/**56.58**/——-/51.06	53.30/45.45/——-/50.80
DIO-top8 + SC	54.05/47.22/——-/58.06	55.35/54.89/——-/57.04	46.13/**51.98**/——-/**52.76**	46.42/55.26/——-/**52.89**	57.33/47.55/——-/51.67
ChOiRe	**54.29/53.33/52.27/58.06**	**60.00/58.77/63.89/58.50**	**52.33**/50.13/**64.76**/51.89	44.74/55.26/64.58/52.76	51.05/**53.74/67.82/53.34**

Table 5: Fine-grained accuracy results of ChatGPT/ChatGPT-Instruct/GPT-4/Mistral. *DIO* stands for *Demographic + Ideology + Opinions* (§4).

Method	ChatGPT
W/o persona	46.60
ChOiRe w/o any personae (Step 4 is used)	**47.79**
DIO-top8 w/o explicit persona	49.22
ChOiRe w/o explicit persona (Steps 2, 3, 4 are used)	**51.55**
DIO-top8 w/o implicit persona	47.16
ChOiRe w/o implicit persona (Steps 1, 4 are used)	**49.97**

Table 6: ChOiRe's generalization results under missing persona(e) situations with ChatGPT. We use Accuracy as the evaluation metric.

Model	ChatGPT	Mistral-7B-Instruct-v0.2
DIO-top8	50.22	44.16
DIO-top8 + FEA	50.64	44.99
DIO-top8 + Random FEA (S=2000)	49.47	42.23
DIO-top8 + Random FEA (S=2024)	48.85	43.36
DIO-LLMtop8	51.03	45.86
DIO + Random LLMtop8 (S=2000)	48.13	44.58
DIO + Random LLMtop8 (S=2024)	49.21	43.84

Table 7: Accuracy results of ChatGPT and Mistral with two trivial variants with two different random seeds 2000 and 2024 in appendix B.2.

and generation outcomes has been studied widely (Xu et al., 2024). For example, Rao et al. (2011); Sakaki et al. (2014) filter the gender information by classifiers; Kim et al. (2017) consider the age while Demszky et al. (2019) analyze the personal political polarity. Our proposed FEA step holds a similar philosophy and can be generalized to and motivative for the above tasks.

- **ChOiRe's Step 2: LLMtop-K.** Selecting top-K most useful individual historical opinions for the next opinion prediction is philosophically related to re-ranking items by LLMs for recommendations (Hou et al., 2024; Xu et al., 2024) and selecting the most utterances in dialogue generation (Do et al., 2022). Undoubtedly, our Step 2 LLMtop-K can be also useful for recommendation tasks.

- **ChOiRe's Step 3: CoO.** Chain-of-Opinion (CoO) is our new innovation from Chain-of-Thought (Wei et al., 2022; Kojima et al., 2022). Essentially, CoO can enhance recommendation tasks by leveraging historical user views in a better way compared to CoT to improve the results and provide explainable recommendations.

- **ChOiRe's Step 4: Majority voting with the dynamic number of historical opinions.** Our step is a creative usage of Self-consistency (Wang et al., 2023b) which is a strong prompting method for LLMs in reasoning tasks. As LLMs are sensitive to selected demonstrations (Perez et al., 2021; Luo et al., 2023; Gao et al., 2023), combining answers from prompting with different historical views can bring more reliable output by then boosting the performance of models. Therefore, this method can potentially be very useful and motivative for recommendation tasks.

In summary, ChOiRe's steps can generalize to and be motivative for other personalized tasks such as recommendation, as either proven or philosophically proven by prior studies.

Model	Accuracy	Collapsed Accuracy
ChatGPT	4.11e-11	6.06e-13
ChatGPT-Inst.	9.97e-8	4.45e-5
GPT-4	4.23e-6	1.17e-9
Mistral	6.01e-8	4.12e-6
GPT-2-large	5.62e-73	6.09e-49
FlanT5-base	1.23e-19	3.19e-12
FlanT5-large	2.55e-21	1.20e-17

Table 8: The p-value computed by student t-test. We observe that all the values are significantly smaller than 0.01 verifying the significance of our improvements.

B.2 Additional Baseline Comparisons

In this section, we compare ChOiRe's FEA and LLMtop-K steps with two simple variants outlined in table 7. Given ChatGPT and Mistral's strong performance with just $4.79/12$ and $8.83/12$ explicit persona attributes, a crucial question arises: *Can comparable performance be achieved by randomly selecting* $5/12$ *and* $9/12$ *explicit persona attributes instead of relying on LLMs?* The first variant, *DIO-top8 + Random FEA*, involves *randomly selecting* $5/12$ and $9/12$ explicit persona attributes. The second variant entails *randomly selecting* 8 implicit persona opinions instead of using ChOiRe's LLMtop-K step. From table 7, we find that randomly selecting explicit persona attributes significantly harms the performance of both models due to the removal of important attributes. Additionally, randomly selecting 8 implicit persona opinions also adversely affects the models, particularly ChatGPT. These observations underscore the effectiveness and importance of ChOiRe's FEA and LLMtop-K steps.

B.3 Student T-test Results for Table 1 & Table 2

We employ the Student t-test to assess the statistical significance between ChOiRe and the best-performing baseline for each model in our study's primary tables: Table 1 and Table 2. Essentially, under the null hypothesis:

- *H0: There is no significant difference.*
- *H1: There is a significant difference.*

The p-values yielded from these tests in Table 8 are remarkably low, well below 0.01.

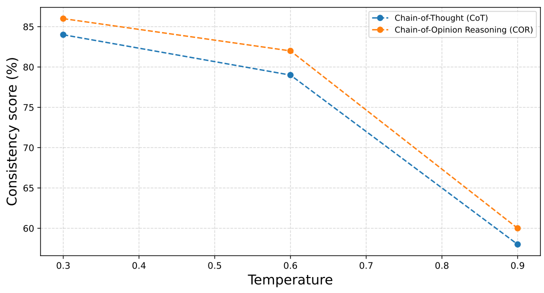

Figure 2: Consistency scores of the baseline DIO-top8 (ChatGPT) with CoO and CoT.

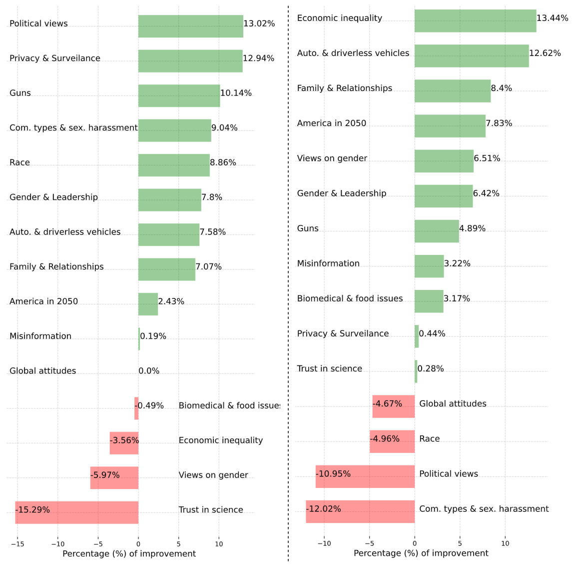

Figure 3: % of improvements over the SOTA method (DIO-top8 + SC) with ChatGPT-Instruct (left) and ChatGPT (right).

B.4 Fine-grained Results of ChatGPT, ChatGPT-Instruct, and GPT-4

Table 5 presents the fine-grained results of ChOiRe and baselines for ChatGPT, ChatGPT-Instruct, and GPT-4.

B.5 Consistency Scores of DIO-top8 with CoO and CoT

Fig. 2 presents the consistency scores of the baseline DIO-top8 (ChatGPT) with CoO and CoT over 100 samples.

B.6 ChOiRe versus Self-Consistency

Fig. 3 presents the improvements per topic comparision between ChOiRe and the SOTA baseline with Self-Consistency (Wang et al., 2023b) DIO-top8 + SC. We observe that ChOiRe improves $11/15$ topics for both ChatGPT and ChatGPT-Instruct.

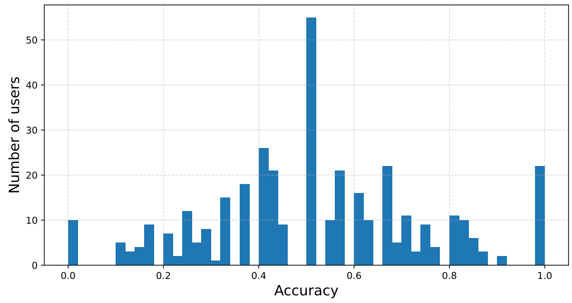

Figure 4: Frequency distribution of accuracy over users by ChOiRe.

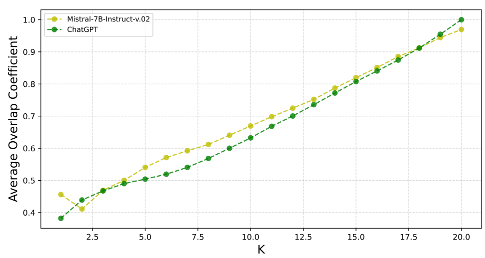

Figure 5: ChatGPT and Mistral-7B-Instruct-v.02 overlap coefficient values for different values of K. We observe that for K is large enough ($K \geq 8$), the coefficient value is relatively acceptable (≥ 0.6).

B.7 Accuracy Distribution over Users

Fig. 4 shows the accuracy distribution over users of ChOiRe with ChatGPT. We see that the peak accuracy is at 0.5 for the majority, with a few users scoring zero and over 20 achieving perfection.

B.8 Top-3 Removed Explicit Personae Attributes

Table 9 presents the top-3 explicit personae that got removed the most by the LLMs. Among the removed personae, `"Citizenship"` appears to be the highest-frequency one across models, followed by `"Race"`.

B.9 Ranking Consistency for LLMtop-K Step

We record the average Overlap coefficient (Vijaymeena and Kavitha, 2016) among 5 ranking outputs from 5 input strategies in fig. 5. The performance of those input strategies is further presented in table 10 on 300 random samples.

B.10 Kendall's Tau Scores for Ranking Agreements

Fig. 6 shows our ranking agreement scores between ChatGPT and Semantic similarity metric (Left), and ChatGPT-Instruct and Semantic similarity metric (Right). We observe that the two ranking orders have minimal monotonous relations with means approximating 0 and low standard deviations. More specifically, with ChatGPT, the maximum agreement is 0.6000 while the minimum is -0.5895 and the Kurtosis is -0.2173. For ChatGPT-Instruct, the maximum is slightly lower with 0.5473, while the minimum is -0.7368 which is smaller ChatGPT, and the Kurtosis is -0.1017.

B.11 Consistency Scores

Table 11 presents the exact consistency scores for the fig. 2. Besides CoO consistently outperforming CoT, we also observe that when the temperature is increased, the consistency score is decreased, which is intuitive.

C Prompts and Prompts Analysis

C.1 Prompt Templates for Filtering Explicit Personae

We present the prompt template for selecting relevant explicit personae for answering the question below. The template is hand-crafted and we use Chain-of-Thought (CoT) prompting (Kojima et al., 2022) via adding `"answer the above question step by step"`.

```
A person can be described by the
following attributes:
{original_attribute_list}
Based on the above list of
demographic information above, now
I give you a new question with
possible answer choices:
Question: '{test_question}'
Answer choices: '{test_choices}'
Please analyze which attributes
in the demographic information are
useful for you to answer the above
question step by step. Give me the
output in the Python list format:
[...]
Give me the answer in the format
below:
Explanations: ...
Answer: [...]
```

C.2 Prompt Templates for Implicit Feature Ranking

We provide our hand-crafted prompt template for ranking implicit personae opinions in the usefulness order below:

Topic	ChatGPT	ChatGPT-Instruct	Mistral-7B-Instruct-v0.2
Guns	'Citizenship', 'Race', 'Marital status'	'Citizenship', 'Frequency of religious attendance', 'Religion'	'Citizenship', 'Education', 'Religion'
Automation & driverless vehicles	'Citizenship', 'Race', 'Marital status'	'Citizenship', 'Race', 'Frequency of religious attendance'	'Citizenship', 'Religion', 'Frequency of religious attendance'
Views on gender	'Citizenship', 'Race', 'Frequency of religious attendance'	'Citizenship', 'Race', 'Frequency of religious attendance'	'Citizenship', 'Religion', 'Frequency of religious attendance'
Community types & sexual harassment	'Citizenship', 'Race', 'Gender'	'Citizenship', 'Frequency of religious attendance', 'Race'	'Education', 'Race', 'Political Party'
Biomedical & food issues	'Citizenship', 'Race', 'Marital status	'Citizenship', 'Race', 'Marital status'	'Citizenship', 'Race', 'Marital status'
Gender & Leadership	'Citizenship', 'Race', 'Region'	'Citizenship', 'Race', 'Frequency of religious attendance'	'Region', 'Race', 'Citizenship'
America in 2050	'Citizenship', 'Race', 'Marital status'	'Citizenship', 'Race', 'Frequency of religious attendance'	'Citizenship', 'Frequency of religious attendance', 'Race'
Trust in science	'Citizenship', 'Marital status', 'Race'	'Citizenship', 'Race', 'Marital status'	'Citizenship', 'Race', 'Region'
Race	'Citizenship', 'Marital status', 'Age'	'Citizenship', 'Age', 'Religion'	'Marital status', 'Education', 'Age'
Misinformation	'Citizenship', 'Marital status', 'Race'	'Citizenship', 'Marital status', 'Race'	'Citizenship', 'Race', 'Religion'
Privacy & Surveillance	'Citizenship', 'Race', 'Marital status'	'Citizenship', 'Race', 'Frequency of religious attendance'	'Religion', 'Race', 'Region'
Family & Relationships	'Citizenship', 'Race', 'Region'	'Citizenship', 'Race', 'Frequency of religious attendance'	'Citizenship', 'Race', 'Religion'
Economic inequality	'Citizenship', 'Frequency of religious attendance', 'Race'	'Citizenship', 'Frequency of religious attendance', 'Race'	'Gender', 'Citizenship', 'Religion'
Global attitudes	'Marital status', 'Race', 'Citizenship'	'Citizenship', 'Marital status', 'Race'	'Gender', 'Frequency of religious attendance', 'Marital status'
Political views	'Citizenship', 'Marital status', 'Frequency of religious attendance'	'Citizenship', 'Frequency of religious attendance', 'Race'	'Frequency of religious attendance', 'Gender', 'Citizenship'

Table 9: Top-3 explicit personae that got removed the most by the LLMs.

Model	Method	Semantic Similarity Order	Seed = 2024	Seed = 5	Seed = 2000	Seed = 15	Std
ChatGPT	DIO-LLMtop8	-	51.03	50.95	51.11	-	0.0652
Mistral-7B-Instruct-v0.2	DIO-LLMtop8	-	45.86	45.55	45.36	-	0.2060

Table 10: Accuracy results of ChatGPT and Mistral on our test set with DIO-LLMtop8 where different orders of input implicit persona opinions are tested for LLMtop-K step.

Model	Temperature	Consistency Score (%)
DIO-top8 + CoT	0.3	84
DIO-top8 + CoO	0.3	86
DIO-top8 + CoT	0.6	79
DIO-top8 + CoO	0.6	82
DIO-top8 + CoT	0.9	58
DIO-top8 + CoO	0.9	60

Table 11: Consistency scores of CoT and CoO on 100 random question-answer pairs. We sample 5 answers per question and measure the % of questions that have all 5 identical answers.

```
Given social behavior
question-answer pairs answered
by a user about his/her opinions
about {subtopic}:
{original_persona_question_order}
You are an expert in analyzing the
social behaviors of a user. Given
a new question asking him/her:
'{test_question}'
Your task is to sort the list
of given question-answer pairs in
descending order such that the
first question-answer pair brings
the most useful information to
answer the new question, whilst the
last question-answer pair brings
the least useful information.
Give me the answer in the form of a
Python list of indexes:
Answer: [...]
```

C.3 Prompt Templates for Baselines Techniques

We use the same prompt templates for ChatGPT (OpenAI, 2022), ChatGPT-Instruct (OpenAI, 2023a), GPT-4 (OpenAI, 2023b). The template prompts for baselines are presented below.

C.3.1 W/o Persona (Santurkar et al., 2023)

The W/o Persona prompt is provided below.

```
Question: {question}
Answer choices:
{choice}
Complete the answer by the
following format without any
explanation:
Answer: A. or B. or C. or D. or E...
```

C.3.2 DIO-top8 (Hwang et al., 2023)

The DIO-top8 prompt is provided below.

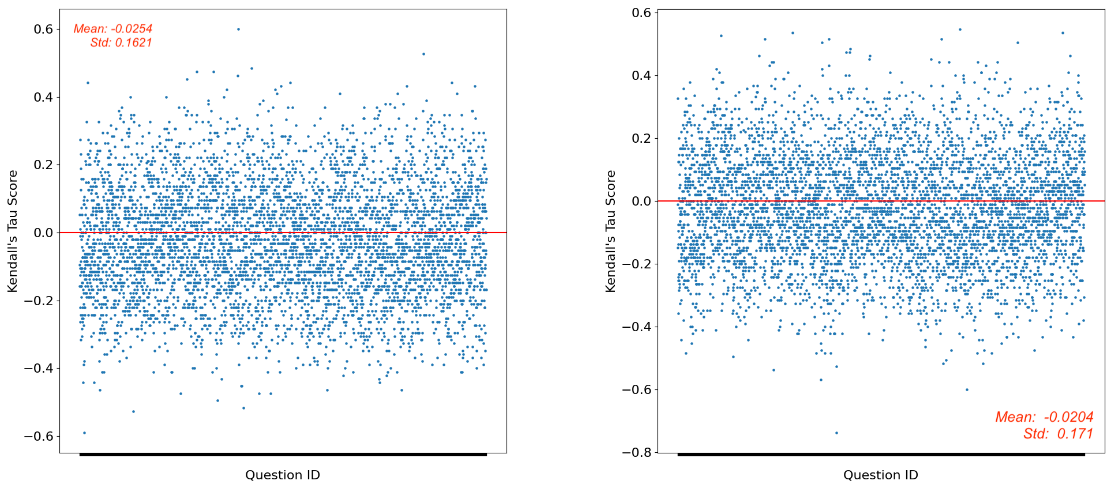

Figure 6: *Left:* Ranking agreements between ChatGPT top-K and semantic similarity top-K. *Right:* Between ChatGPT-Instruct top-K and semantic similarity top-K. One example that has a high disagreement score is shown in appendix E.2.

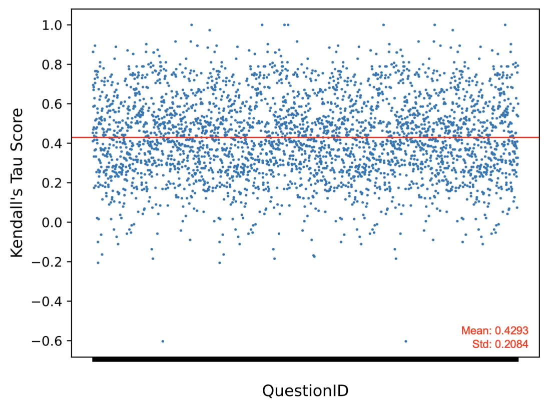

Figure 7: Ranking agreements between Mistral top-K and semantic similarity top-K.

```
A person can be described as follows:
{explicit_persona_str}
The person has the following opinions on {topic}.
Opinions:
{implicit_persona_str}
Based on the above information, which answer choice is the user most likely to choose?
Question: {question}
Answer choices: {choice}
Give the answer in the format:
Answer: A. or B. or C. or D. or E....
```

C.3.3 Self-refine (Madaan et al., 2023)

The Self-refine prompts (Madaan et al., 2023) are provided below, feedback step and refine step respectively.

```
You are given a question and an answer for that question. Analyze the question and the answer and provide some feedback on the answer to the question. Don't change the answer, just provide feedback.
Question: {test_question}
Choices: {choices}
Answer: {selected_choice}
Feedback:
```

```
You are given a question, an answer to that question and a feedback to the answer. Based on the feedback, refine your answer and generate the final answer in around 170 words.
Question: {test_question}
Answer: {selected_choice}
Feedback: {feedback}
Refined answer: new_choice + explanation
```

C.3.4 Chain-of-Thought (Kojima et al., 2022)

The CoT prompt template is provided below.

```
A person can be described as
follows:
{explicit_persona_str}
The person has the following
opinions on {topic}.
Opinions:
{implicit_persona_str}
Based on the above information,
answer the following question
step-by-step:
Question: {question}
Answer choices: {choice}
Give the answer in the format:
Answer: A. or B. or C. or D. or
E....
Explanations:...
```

C.3.5 Chain-of-Opinion (Ours)

Our CoO prompt template is provided below.

```
A person can be described as
follows:
{explicit_persona_str}
The person has the following
opinions on {topic}.
Opinions:
{implicit_persona_str}
Based on the above information,
answer the following question
step-by-step by explaining and
analyzing each of the Opinions and
Demographic Information:
Question: {question}
Answer choices: {choice}
Give the answer in the format:
Answer: A. or B. or C. or D. or
E....
Explanations:...
```

C.4 Prompting Costs for API Models

Our prompting costs for API models are reported in table 12. We observe that for GPT-4, ChOiRe costs a similar price with the baseline DIO-top8 while DIO-top8 + SC costs nearly double our price. This is because we perform the FEA and LLMtop-K steps of ChOiRe by ChatGPT, which are relatively cheap. Additionally, for ChatGPT and ChatGPT-Instruct, ChOiRe costs around 7 and 10 more US$ dollars in total compared to baseline DIO-top8 + SC. However, these extra amounts of costs are worth it because we gain significant improvements over all the baselines and especially huge improvements for some topics.

D Human Evaluation

D.1 Human Rating Instructions

Our details of human rating instructions are provided in table 13 for all the criteria. It is worth noting that selecting all features can't get a high FEA Satisfaction score, according to our instructions. In addition, if the selected explicit personae fall among several scores, the annotators are instructed to take the minimum score.

E Examples

E.1 FEA Example with ChatGPT

Fig. 8 shows an FEA example with ChatGPT. We observe that by removing unnecessary explicit personae including `"Age"`, `"Citizenship"`, `"Education"`, `"Income"`, `"Marital Status"`, `"Race"`, `"Frequency of religious attendance"`, ChatGPT predicts the opinion accurately, while without removing, an incorrect prediction was made.

E.2 Example of High Disagreement between Rankings

Fig. 9 illustrates one example of the high disagreement between orders by semantic similarity scores and LLM (ChatGPT). We derive three observations, as discussed in §6.1. First, not all top-8 opinions by semantic similarity scores help predict the opinion. For example, 16-th opinion, despite having a relatively high semantic similarity score with the question which might offer some perspective on the prevalence of guns in the user's community during the upbringing, is less directly related to the question. This is similar to the 18-th opinion which is also less relevant. Meanwhile, several important opinions are deselected by the semantic-similarity-based method, such as the $6, 3, 4, 10$-th ones, which are chosen by the LLM. The 6-th one is critical, and directly relevant because it assesses the person's attitude toward safety measures related to gun ownership. Finally, by using LLMtop-K order, the model predicts the opinion accurately, while an incorrect prediction is made with the semantic similarity order.

	DIO-top8	DIO-top8 + CoT	DIO-top8 + SC	ChOiRe	Model
Ave. consumed #tokens	562.72	623.62	995.89	3142.86	ChatGPT
Total US$	3.01	3.73	6.82	13.95	ChatGPT
Ave. consumed #tokens	562.72	630.58	1019.31	3121.72	ChatGPT-Instruct
Total US$	3.12	3.84	7.11	19.99	ChatGPT-Instruct
Ave. consumed #tokens	559.27	-	1021.14*	3180.82	GPT-4
Total US$	91.19	-	226.15*	123.30	GPT-4

Table 12: Prompting cost analysis of ChOiRe and other baselines as of 1st Feb 2024. * denotes our estimation on 50 random samples.

Criterion	**Scoring Instruction**
FEA Satisfaction	1: The number of filtered-out explicit personae that are directly relevant for answering the question is more than 3. 1: The number of selected explicit personae that are somewhat irrelevant for answering the question is more than 3. 2: The number of filtered-out explicit personae that are directly relevant for answering the question is 2 or 3. 2: The number of selected explicit personae that are somewhat irrelevant for answering the question is 2 or 3. 3: The number of filtered-out explicit personae that are directly relevant for answering the question is less than or equal to 1. 3: The number of selected explicit personae that are somewhat irrelevant for answering the question is less than 2.
LLMtop-K Satisfaction	1: Among the top-8 implicit persona opinions, the number of less relevant opinions for answering the question is more than 4. 2: Among the top-8 implicit persona opinions, the number of less relevant opinions for answering the question from 2 to 4. 3: Among the top-8 implicit persona opinions, the number of less relevant opinions for answering the question is less than or equal to 1.
CoO Reasonableness	1: The CoO has limited or flawed reasoning thoughts with inadequate support. 2: The CoO has some reasoning thoughts with decent support but room for improvement. 3: The CoO has strong, clear, and well-supported reasoning thoughts with a comprehensive understanding.
CoO Follow the Instruction	1: The generated CoO explanation does not mention more than 4 attributes/opinions from explicit and implicit personae. 2: The generated CoO explanation somewhat follows the instruction by involving more than 4 attributes/opinions but room for improvement. 3: The generated CoO explanation follows perfectly the instruction via explaining all the explicit and implicit attributes one by one.

Table 13: Human rating instructions. FEA, LLMtop-K, and CoO stand for Filtering Explicit Personae Attributes, Implicit Personae Opinions Ranking, and Chain-of-Opinion reasoning (§3).

E.3 Example of Inconsistent Answers Generated by CoT

Fig. 10 illustrates an example of the inconsistent answers generated by ChatGPT with Chain-of-Thought (Kojima et al., 2022) (CoT). It is observed that different subsets of top-8 implicit personae opinions are mentioned in the two explanations, leading to varied final answers.

E.4 Example of Chain of Opinion Reasoning

Fig. 11 presents an example of the answer generated by ChatGPT using Chain of Opinion (ours) versus Chain of Thought (Wei et al., 2022) prompting methods.

E.5 Example of Answer Consistency with Dynamic Numbers of Opinions

Fig. 12 shows an example of the answer generated by GPT-4 using Chain of Opinion (ours) reasoning with different numbers of provided historical opinions.

E.6 Error Analysis Examples

Fig. 13 illustrates our error analysis examples of ChOiRe with ChatGPT. The top-left frame is an example of FEA missing key explicit personae. The bottom one is an instance demonstrating the error of the LLMtop-K algorithm including less relevant opinions. The top-right rectangular is an example from GPT-4, showing that it does not follow human instructions to predict opinion via chain-of-opinion reasoning.

DIO-top8

A person can be described as follows:

Age: 65+
Citizenship: Yes
Region: South
Education: Some college, no degree
Income: $30,000-$50,000
Marital status: Married
Political ideology: Very conservative
Political party: Republican
Race: White
Religion: Protestant
Frequency of religious attendance: More than once a week
Gender: Male

The person has the following opinions on Guns:
...
Question: How much, if at all, do you think family instability contributes to gun violence in the country today?

Answer choices:
A. A great deal, B. A fair amount, C. Not too much, D. Not at all

Answer: D. Not at all ✗

DIO-top8 + FEA

A person can be described as follows:

Region: South
Political ideology: Very conservative
Political party: Republican
Religion: Protestant
Gender: Male

The person has the following opinions on Guns:
...
Question: How much, if at all, do you think family instability contributes to gun violence in the country today?

Answer choices:
A. A great deal, B. A fair amount, C. Not too much, D. Not at all

Answer: C. Not too much ✓

Figure 8: FEA example with ChatGPT.

Question: Would having a gun in your household make you feel?

Answer choices:

A. Safer than you feel without a gun in your household

B. Less safe than you feel without a gun in your household

C. No more or less safe

0. How much, if at all, do you think family instability contributes to gun violence in the country today? (**answer:** A fair amount)
1. Do you feel that people in your local community tend to look at most gun owners in a positive way or a negative way? (**answer:** Negative way)
2. How much, if at all, do you worry about the following happening to you? Having a personal health crisis (**answer:** Worry a lot)
3. How much, if at all, do you think the ease with which people can illegally obtain guns contributes to gun violence in the country today? (**answer:** A great deal)
4. Would you say the following was a reason or was not a reason why there were guns in your household when you were growing up? For sport shooting, including target shooting and trap and skeet (**answer:** No, was not a reason)
5. How often, if ever, do you watch TV programs about guns or watch gun-oriented videos (**answer:** Never)
6. Thinking about gun owners who have children in their home, how important do you think it is for them to: Take gun safety courses (**answer:** Important but not essential)
7. How often, if ever, do you go shooting or to a gun range? (**answer:** Never)
8. How safe, if at all, would you say your local community is from crime? Would you say it is (**answer:** Somewhat safe)
9. As far as you know, how many of your friends, if any, own guns? (**answer:** None)
10. Thinking about people who commit suicide using a gun, which comes closer to your view, even if neither is exactly right? (**answer:** They would find a way to do it whether they had access to a gun or not)
11. Do you personally own any guns (not including air guns, such as paintball, BB or pellet guns)? (**answer:** No, I don't own any guns)
12. Do you feel that society in general tends to look at most gun owners in a positive way or a negative way? (**answer:** Negative way)
13. How much, if at all, do you worry about the following happening to you? Not being able to pay your bills (**answer:** Worry a little)
14. Thinking about when you were growing up, as far as you know, were there ever any guns in your household or not? (**answer:** Yes, there were guns in my household)
15. Does anyone else in your household own any guns (not including air guns, such as paintball, BB or pellet guns)? (**answer:** No, no one else in my household owns a gun)
16. Thinking about the people in the community where you spent the majority of time when you were growing up, as far as you know, how many people owned guns? (answer: Only a few)
17. Regardless of whether or not you own a gun, have you ever fired a gun? (**answer:** Yes, I have fired a gun)
18. Would you say the following was a reason or was not a reason why there were guns in your household when you were growing up? For hunting (**answer:** No, was not a reason)
19. Thinking about gun owners who have children in their home, how important do you think it is for them to: Keep all of their guns unloaded (**answer:** Essential)

Semantic similarity order: [*12, 14, 19, 17, 15, 18, 16, 9,* 1, 0, 6, 10, 11, 4, 8, 3, 7, 5, 13, 2] ===> **ChatGPT answer:** A. Safer than you feel without a gun in your household ✗

LLM (ChatGPT) order: [6*, 15, 14, 4, 3, 9, 10, 12,* 17, 16, 18, 1, 5, 7, 8, 11, 13, 0,2, 19] ===> **ChatGPT answer:** C. No more or less safe ✓

Figure 9: Example of the high disagreement between orders by semantic similarity scores and LLM (ChatGPT).

A person can be described as follows:

Age: 65+
Citizenship: Yes
Region: West
Education: College graduate/some postgrad
Income: Less than $30,000
Marital status: Widowed
Political ideology: Moderate
Political party: Independent
Race: Asian
Religion: Roman Catholic
Frequency of religious attendance: Once a week
Gender: Male

The person has the following opinions on Guns.

1. How much, if at all, do you think the ease with which people can illegally obtain guns contributes to gun violence in the country today? (answer: A great deal)
2. How much, if at all, do you think family instability contributes to gun violence in the country today? (answer: A fair amount)
3. Do you feel that people in your local community tend to look at most gun owners in a positive way or a negative way? (answer: Negative way)
4. How much, if at all, do you worry about the following happening to you? Having a personal health crisis (answer: Worry a lot)
5. Would you say the following was a reason or was not a reason why there were guns in your household when you were growing up? For sport shooting, including target shooting and trap and skeet (answer: No, was not a reason)
6. How often, if ever, do you watch TV programs about guns or watch gun-oriented videos (answer: Never)
7. Thinking about gun owners who have children in their home, how important do you think it is for them to: Take gun safety courses (answer: Important but not essential)
8. How often, if ever, do you go shooting or to a gun range? (answer: Never)

Based on the above information, answer the following question step-by-step:

Question: How much, if at all, do you think the amount of gun violence in video games contributes to gun violence in the country today?
Answer choices:
A. A great deal
B. A fair amount
C. Not too much
D. Not at all

Give the answer in the format:
Answer: A. or B. or C. or D. or E....
Explanations:...

Response 1

Answer: D. Not at all
Explanations: The person's opinions on guns do not provide any indication of their beliefs about the contribution of gun violence in video games to gun violence in the country.

Response 2

Answer: C. Not too much
Explanations: The person's opinions on guns do not indicate any concern or belief that gun violence in video games contributes significantly to gun violence in the country. They did not mention it as a reason for guns in their household when growing up, they do not watch TV programs about guns or gun-oriented videos, and they did not express any worry about the amount of gun violence in video games. Therefore, it can be inferred that they believe the amount of gun violence in video games does not contribute much to gun violence in the country today.

Figure 10: Example of the inconsistent answers generated by ChatGPT with Chain-of-Thought.

Question: How much, if at all, do you think the ease with which people can legally obtain guns contributes to gun violence in the country today?
Answer choices: A. A great deal; B. A fair amount; C.Not too much; D.Not at all

Chain of Opinion Reasoning ✓
To answer the question about how much the ease of legally obtaining guns contributes to gun violence in the country today, we need to consider the individual's demographic information and opinions. Let's analyze each of the provided opinions and demographics and then provide an answer based on the available information:
…
Opinion 1: The individual believes that family instability contributes a fair amount to gun violence.
Opinion 2: The individual thinks that the amount of gun violence in video games contributes a fair amount to gun violence.
…
Based on this information, **it is clear that the individual holds opinions suggesting they may perceive a significant relationship between access to guns and gun violence**. They express concerns about gun violence in various contexts, such as family instability and video games, and emphasize the importance of gun safety. Given these opinions, it is likely that they would choose **answer choice A. A great deal**.

Chain of Thought ✗
The person's demographic information suggests that they are a liberal, Democrat, with a postgraduate education, and a Roman Catholic faith. They are living with a partner, belong to the West region, and have a relatively low income. They are also a female aged 50-64 and do not frequently attend religious services. The person's opinions indicate that they are concerned about various aspects of gun violence, such as family instability, the influence of gun violence in video games, and the importance of gun safety for gun owners with children in their homes. However, they do not worry about being the victim of a mass shooting. Based on this analysis, it is likely that this individual would believe that the ease with which people can legally obtain guns contributes at least a fair amount to gun violence in the country today. **Answer: B. A fair amount**

Figure 11: Example of an answer generated by Chain of Opinion versus Chain of Thought prompting with ChatGPT.

Chain of Opinion Reasoning (K=8)

...

The given opinions and demographic information do not provide any insight into the person's financial situation or their concerns about financial stability. **Therefore, the answer cannot be determined.**

Chain of Opinion Reasoning (K=10)

...

Given these demographic and opinion insights, it can be inferred that this person may worry a little (option B) about not being able to pay their bills. While their income is relatively low, their concerns about personal well-being and their awareness of societal issues related to violence suggest that they may have some level of concern about financial stability. **Answer: B. Worry a little**

Chain of Opinion Reasoning (K=12)

...

Given the information provided, it is likely that the person may have some financial worries due to their lower income, but they also have concerns about personal safety. Their opinions on gun-related issues and safety measures for gun owners with children indicate that they value safety and responsible gun ownership. **Answer: B. Worry a little**

Figure 12: Example of our answer consistency technique (ours), generated by GPT-4.

Question: For each one of the following, please indicate whether you think it is a reason why there aren't more women in top executive business positions. Sexual harassment creates an environment that makes it harder for women to succeed in business.
Answer: Major reason
ChatGPT Attributes: "Gender", "Marital status"

Annotators' FEA Satisfaction: 2
Reason: missing "Education" and "Age"

...
Based on the above information, answer the following question step-by-step by explaining and analyzing each of the Opinions and Demographic Information:
Question: How often, if ever, do you listen to gun-oriented podcasts or radio?
...
GPT-4 Answer: C. Hardly ever
Explanations: This person is a 65+ male from the South, identifies as very conservative, and is a member of the Republican party. These demographic factors might suggest a higher likelihood of interest in guns and gun-related activities. However, his responses to the questions about his engagement with gun-related activities and media suggest otherwise...

Annotators' Follow the Instruction: 1
Annotators' Reasonableness: 3
Reason: Do not follow the instruction.

Question: Do you think it's a good thing or a bad thing that most people in our society look up to women who are womanly or feminine?
Answer: Good thing
LLMtop-8 ChatGPT:
1. Thinking about how society sees women these days, in general, would you say (**answer:** Most people look up to women who are womanly or feminine)
2. Which statement comes closer to your views, even if neither is exactly right? (**answer:** Whether someone is a man or a woman is determined by the sex they were assigned at birth)
3. Which of the following statements comes closer to your feelings? (**answer:** Our society has been about right when it comes to accepting people who are transgender)
4. Do you now smoke cigarettes (**answer:** Not at all)
5. Thinking about your work life, has increased outsourcing of jobs to other countries done more to help or hurt your job or career, or hasn't it made much of a difference for your job or career? (**answer:** Hasn't made much of a difference)
6. Thinking about your work life, has automation of jobs through new technology in the workplace done more to help or hurt your job or career, or hasn't it made much of a difference for your job or career? (**answer:** Hasn't made much of a difference)
7. Thinking about your work life, has a growing emphasis on diversity in the workplace done more to help or hurt your job or career, or hasn't it made much of a difference for your job or career? (**answer:** Hasn't made much of a difference)
8. Thinking about your work life, has more foreign-made products being sold in the U.S. done more to help or hurt your job or career, or hasn't it made much of a difference for your job or career? (**answer:** Has done more to help my job or career)

Annotators' Ranking Satisfaction: 1
Reason: Opinions 4th, 5th, 6th, 7th, 8th are less relevant. Meanwhile, other opinions below are more relevant, for example:
9. In general, how much pressure, if any, do you think men face in our country these days to join in when other men are talking about women in a sexual way? (**answer:** Not too much)
10. When it comes to raising girls, would you say there is too much emphasis or too little emphasis on encouraging girls to do well in school these days, or is it about right? (**answer:** About right)

Figure 13: Error analysis examples of ChOiRe with ChatGPT.