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.*

Aligning language models (LMs) with human opinion is challenging yet vital to enhance their grasp of human values, preferences, and beliefs. We present ChOiRe , a four-step solution framework to predict human opinion that differentiates between the user's *explicit personae* (i.e. demographic or ideological attributes) that are manually declared, and *implicit personae* inferred from user historical opinions. Specifically, it 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 LLM-based techniques significantly by 3.22%.

1 Introduction

With the growing volume of human–AI interactions, language models (LMs) are emerging as powerful supportive tools such as dialogue agents (OpenAI, 2022; Google, 2022), writing assistants (Wordtune, 2022) and decision-making supporters (Ye et al., 2023). Aligning them with users' unique personalities[1] — personalization — is crucial for meeting individuals' expectations and delivering tailored experiences.

Recent personalization research with LMs has emphasized aligning them with user groups using persona-based prompts (Santurkar et al., 2023; Deshpande et al., 2023; Argyle et al., 2023). However, LMs form opinions based on their training data and feedback, resulting in low steerability even for well-represented user groups (Santurkar et al., 2023). This raises concerns about their effectiveness in steering for individual users. Furthermore, individuals hold nuanced opinions that evolve over time and are influenced by situational factors. These observations identify the challenge in aligning LMs with individuals, which remains much less explored.

Recently, Hwang et al. (2023) found significant opinion variations among individuals sharing the same demographics, exposing flaws in current group-focused LMs alignment. They argue for individualised models, introducing an approach integrating a user's demographic & ideological attributes, (which we term as *explicit personae*) and user historical opinions (*implicit personae*) into the prompt context for opinion prediction.

While this naïve strategy achieves good results, we argue that it suffers from a few key limitations. First, it employs all explicit personae. However, we contend that only a subset is necessary for accurate opinion prediction, and including non-relevant personae may act as noise, harming predictive performance (Appendix A.5). Second, Hwang et al. (2023) utilize the top-K semantically similar opinions with respect to the question (here termed top-K implicit personae). Our argument is that this approach is inefficient, as the opinions ranked highest in semantic similarity may not offer the most valuable information for opinion prediction (§6.1). Our empirical experiments suggest that LMs may lack sufficient personae evidence with a fixed K (Table 4) — dynamically adjusting K per task can overcome such deficiencies. Finally, while Chain-of-Thought (CoT; Wei et al. 2022) enables LMs to explicate intermediate reasoning steps to perform multi-step reasoning tasks effectively, we find that naïve application of CoT does not help this task

[1] Except where otherwise specified, we use the terms "human", "individual" and "user" interchangeably.

(§5). We consider re-engineering CoT for opinion analysis.

To address the above challenges, we propose ChOiRe[2], a four-step solution framework for opinion prediction. First, a large language model (LLM) is employed to analyze the user's explicit personae to discard irrelevant ones. Second, the LLM ranks implicit persona opinions in usefulness order and selects the top-K as the most valuable, surpassing 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. ChOiRe applies self-consistency over CoO to provision the right amount of personal information for opinion inference. ChOiRe achieves new state-of-the-art (SOTA) effectiveness while sparingly using limited inference calls. In detail, our contributions are:

1. We highlight the importance and limitations of prior opinion prediction work. We propose ChOiRe as a four-step framework overcoming these limitations;

2. ChOiRe surpasses prior efforts significantly, achieves SOTA results with limited inference calls using modern LLMs, enhancing models' reliability in predicting human opinions;

3. We conduct a thorough analysis to verify our hypotheses concerning explicit and implicit personae and chain-of-opinion reasoning.

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. 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. We argue that this naïve strategy is suboptimal as discussed in §1, and 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 (§5), but that an appropriate modification does. We propose Chain-of-Opinion (CoO) reasoning (§3) that overcomes CoT's limitations in this task. While we note other new 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 debuted, 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 notated as $D = \{\langle T, E, I, q, a\rangle_n\}_{n=1}^{N}$, where T, E and I indicate the topic of a question q, the *explicit personae* and *implicit personae* of the user-answered q, and a is the q's answer. 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 the user's personae and predict the opinion a, given T, E, I, q.

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 per-

[2] **Ch**ain of **Opi**nion **Re**asoning, pronounced as the English word "choir".

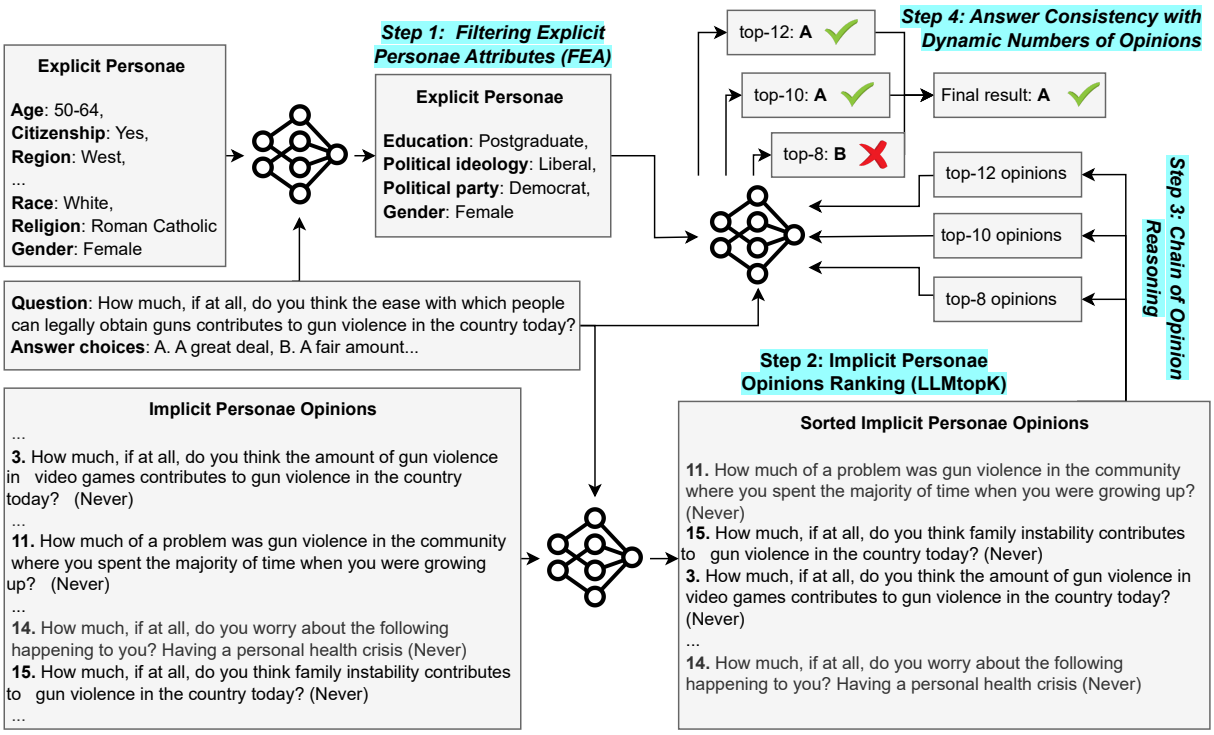


Figure 1: ChOiRe overview, consisting of the four main steps (cyan background), as detailed in §3.

sonae, 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 include the topic information T in all the prompts. We use one LLM for all the steps, except GPT-4 (OpenAI, 2023b) due to its high computational expenses. We present the details of each step 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"`, `"income"`, and `"political ideology"` — has been shown to help models characterize and predict human opinions more accurately (Hwang et al., 2023). However, *which personae matters and which do not?* is still an open question. While Hwang et al. (2023) use all these attributes to guide the language model in predicting the opinions, we argue that not all of them are necessary for the model to accurately predict the opinions, and they may even harm its predictions. Appendix A.5 shows such an example where with all the explicit personae, the model made a wrong prediction while removing unnecessary personae the model made a correct prediction. Whilst we may assume that LLMs can self-ignore irrelevant information, the performance change can be explained by the fact that LLMs use the attention mechanism (Bahdanau et al., 2015; Luong et al., 2015), that attends to all the tokens in the input prompts, and the attentions can be imperfect in neglecting irrelevant tokens. To address this problem, we propose a simple strategy to filter out unnecessary explicit personae. Specifically, an LLM is employed to analyze how each persona is helpful for the model to predict the opinion via Chain-of-Thought (Wei et al., 2022). The model then outputs a list of helpful personae, given the question and the opinion answer choices. We then use only the selected attributes for predicting the opinions. The prompt template is provided in Appendix A.1. Surprisingly, we find that LLMs evaluate more than half of the explicit personae as not useful 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 different orders of demonstrations (Perez et al., 2021; Luo et al., 2023; Gao et al., 2023). In the context of predicting human opinions, we discover that LLMs are also sensitive to the chosen implicit personae opinions, and their orders in the prompts. Hwang et al. (2023) rank the implicit personae opinions via semantic-similarity scores and selects top-K. We argue that this strategy is sub-optimal 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 A.6). To address this challenge, we propose to utilize LLMs to analyze and rank the implicit personae opinions in usefulness descending order instead. The prompt template is illustrated in Appendix A.2. By doing so, our proposed method supports the usefulness in predicting the opinions in terms of meanings, rather than the semantic similarity. We name 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., 2023). Furthermore, it is little known how the models perceive multiple implicit personae opinions, especially when many opinions are provided, *which one(s) the models used, which one(s) they didn't for predicting the opinion?* Our preliminary experiments with CoT (§6.1 and appendix A.8) 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[3] (see Appendix A.7). 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 main advantages despite its simplicity. 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.

[3] greater or equal to 0.6

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 A.9). We attribute this to insufficient user implicit personae opinions provided. Inspired by *Self-Consistency (SC)* (Wang et al., 2023), our approach involves sampling multiple answers using different K values for a given question. The most frequent answer, along with its explanation, 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.

4 Evaluation

Dataset. We experiment on OpinionQA dataset (Santurkar et al., 2023), a benchmark 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.

Baseline Models. We use ChatGPT (OpenAI, 2022), ChatGPT-Instruct (OpenAI, 2023a), as our LLMs, and compare ChOiRe with 4 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 (Wei 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., 2023) to *DIO-top8* to select the most frequent answer generated by the model as the final opinion prediction. It is worth noting that we do not conduct the experiment with InstructGPT (Ouyang et al., 2022) like Hwang et al. (2023) since this model is going to be deprecated (OpenAI, 2023a) and replaced by ChatGPT-Instruct. We also run three main experiments with GPT-4 (OpenAI, 2023b) to verify the effectiveness of our proposed framework. However for GPT-4, due to the budget limit, we use ChatGPT for FEA and LLMtop-K steps. The full prompts used for the models are presented in Appendix A.3.

Metrics. We employ Accuracy as the primary metric for automatic evaluations, following Hwang et al. (2023). Additionally, human evaluations are crucial due to the absence of automated metrics assessing LLMs' performance in intermediate steps of ChOiRe. Therefore, we conduct our human assessments to address these research questions: *(1) LLMs' effectiveness in filtering unecessary 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. We then hire 3 annotators who are English native speakers. For filtering explicit personae and ranking implicit opinions steps, each annotator is instructed to rate on a 1-3 scale (3 is the 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 inter-annotators' agreement is assessed by Krippendorff's alpha (Krippendorff, 2011). Our human rating instructions are provided in Appendix A.13.

Model	ChatGPT	ChatGPT-Inst	GPT-4
W/o persona	46.60	44.91	-
DIO-top8	50.22	51.95	57.98
DIO-top8 + CoT	49.96	51.90	-
DIO-top8 + SC	50.58	52.06	-
DIO-top8 + FEA	50.64	52.63	-
DIO-top8 + CoO	50.97	52.08	-
DIO-LLMtop8	51.03	52.80	-
DIO-LLMtop8 + FEA	51.19	52.97	-
DIO-LLMtop8 + FEA + CoO	51.90	53.01	59.02
ChOiRe	**52.21**	**53.26**	**59.30**
% Improvements	+3.22	+2.52	+2.28

Table 1: Overall accuracy on ChatGPT, ChatGPT-Instruct, and GPT-4. *FEA* is our first step, stands for *Filtering Explicit Attributes*. *LLMtop8* the second step, ranking the implicit persona opinions via LLM, and selecting top-8 as the input, and *CoO* stands for *Chain-of-Opinion reasoning*.

Implementation Details. 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. We use Nucleus Sampling (Holtzman et al., 2020) with a $p = .9$ 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 technique for each question.

5 Main Results

Automatic Evaluation Results. Table 1 shows our main experimental outcomes and Table 2 illustrates the fine-grained results among the benchmark topics. We derive six main observations. First, among the baselines in Table 1, naïve CoT (via simply adding `"answer the following question step-by-step"`) slightly reduces the performance of models with *DIO-top8* (Hwang et al., 2023) in this task, while SC improves marginally. We attribute this to the inconsistency of CoT explanations (§3). Second, compared with (Hwang et al., 2023), ChOiRe improves the performance of ChatGPT, ChatGPT-Instruct, and GPT-4 significantly

Model	ChatGPT/ChatGPT-Inst/GPT-4				
	Guns	**Auto. & driverless vehicles**	**Views on gender**	**Com. types & sex. harassment**	**Race**
W/o persona	53.07/37.30/-	47.73/48.26/-	50.53/42.94/-	**47.73**/41.67/-	41.95/45.28/-
DIO-top8	53.87/57.00/60.39	45.33/44.78/**53.22**	53.21/52.15/63.73	43.47/**45.24**/42.86	43.06/44.65/**55.17**
DIO-top8 + CoT	54.55/52.33/-	47.22/46.77/-	48.11/**57.67**/-	42.39/42.26/-	**45.63**/43.40/-
DIO-top8 + SC	54.40/52.85/-	43.73/48.26/-	55.61/56.44/-	45.33/40.48/-	45.00/43.40/-
ChOiRe	**57.06/58.21/63.37**	**49.25/51.92**/50.00	**59.23**/53.07/**71.43**	39.88/44.14/**47.96**	42.77/**47.28**/50.57
	Gender & Leadership	**America in 2050**	**Trust in science**	**Biomedical & food issues**	**Misinformation**
W/o persona	**53.13**/50.83/-	39.73/39.13/-	50.40/47.29/-	53.87/53.63/-	46.93/40.38/-
DIO-top8	48.27/54.70/**65.55**	46.93/46.20/43.70	54.93/61.58/61.54	52.27/55.86/58.03	49.33/52.11/52.71
DIO-top8 + CoT	48.58/50.83/-	43.05/48.91/-	54.10/65.02/-	**56.91/57.54**/-	**49.57/53.99**/-
DIO-top8 + SC	49.07/53.60/-	45.87/47.83/-	56.27/**65.52**/-	53.07/**57.54**/-	45.00/53.52/-
ChOiRe	52.22/**57.78**/63.03	**49.46/48.99/45.37**	**56.43**/55.50/**68.46**	54.75/57.26/**61.61**	46.45/53.62/**57.36**
	Privacy & Surveilance	**Family & Relationships**	**Economic inequality**	**Global attitudes**	**Political views**
W/o persona	43.24/40.28/-	47.06/44.36/-	43.67/49.15/-	46.13/46.71/-	40.80/48.95/-
DIO-top8	53.24/47.22/47.73	57.22/57.89/62.50	45.60/**51.98**/63.81	**49.60**/57.23/**66.67**	**56.80**/46.85/62.07
DIO-top8 + CoT	53.38/47.22/-	59.57/55.64/-	47.65/**51.98**/-	46.42/**56.58**/-	53.30/45.45/-
DIO-top8 + SC	54.05/47.22/-	55.35/54.89/-	46.13/**51.98**/-	46.42/55.26/-	57.33/47.55/-
ChOiRe	**54.29/53.33/52.27**	**60.00/58.77/63.89**	**52.33**/50.13/**64.76**	44.74/55.26/64.58	51.05/**53.74/67.82**

Table 2: Fine-grained accuracy results of ChatGPT/ChatGPT-Instruct/GPT-4. *DIO* stands for *Demographic* + *Ideology* + *Opinions* (§4).

with 3.22%, 2.52% and 2.28% accuracy. It establishes a strong SOTA result with GPT-4, surpassing InstructGPT (Ouyang et al., 2022) by a notable margin of 53.74% in our testing set. 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. Third, diving deeper into the benchmark topics in Table 2, ChOiRe achieves SOTA results in $8/15$ topics for ChatGPT and ChatGPT-Instruct, notably improving in *Economic Inequality* and *Privacy & Surveillance*. Additionally, GPT-4 attains the best results in $11/15$ topics. Fourth, comparing with the best baseline DIO-top8 + SC, Fig. 2 illustrates the percentage of improvements gained from ChOiRe with ChatGPT and ChatGPT-Instruct. We observe that ChatGPT-Instruct yields the most significant improvement in the *Political views* category, while it experiences the biggest performance drop in *Trust in science*. ChatGPT excels in enhancing and experiencing declined performance in the *Economic inequality* and *Community types and sexual harassment* topics, respectively. Fifth, Fig. 3-Left illustrates the accuracy distribution over the users of ChOiRe with ChatGPT. We see that the model's peak accuracy is at 0.5 for the majority, with a few users scoring zero and over 20 achieving perfection. Finally, ChatGPT (OpenAI, 2022) and ChatGPT-Instruct (OpenAI, 2023a) show improvements by selecting only $4.79/12$ and $5.59/12$ explicit personae on average, respectively. This suggests that over half of explicit personae may not contribute significantly to opinion prediction.

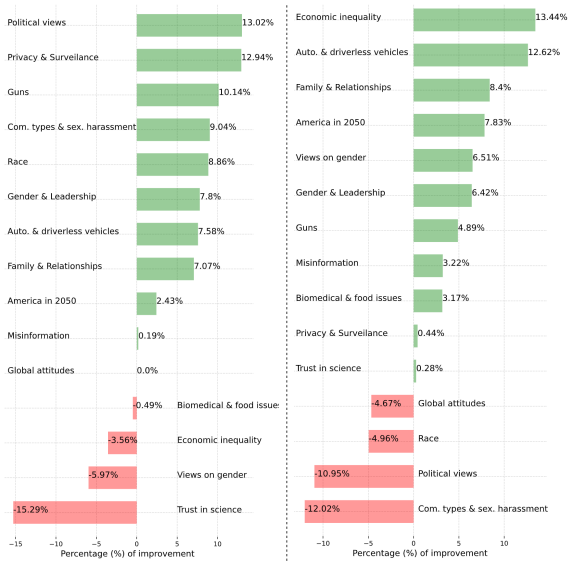


Figure 2: % of improvements over the SOTA method (DIO-top8 + SC) with ChatGPT-Instruct (left) and ChatGPT (right).

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)

Table 3: Human evaluation results. Kα' represents the Krippendorff's alpha score.

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 ChatGPT excels slightly in ranking, ChatGPT-Instruct performs slightly better in explicit personae selection. Both 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, three models effectively generate intermediate reasoning 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, 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

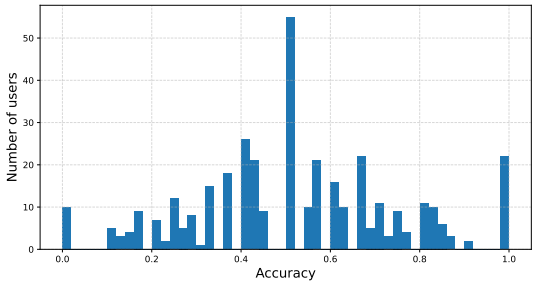
6.1 Methodology Analysis

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% and 1.3% performance boost on ChatGPT and ChatGPT-Instruct, 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 language models (LLMs) across various topics, we document the top-3 removed personae in Appendix A.4. 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, while ChatGPT-Instruct commonly removes `"Frequency of religious attendance"` revealing potential biases in LLMs.

LLMtop-K versus Top-K. We compare the performance of LLMs when being provided the top-K opinions ranked by the LLMs via the usefulness (*DIO-LLMtop8*), and ranked by semantic similarity scores (DIO-top8). From Table 1, DIO-LLMtop8 outperforms DIO-top8 by 1.6% on both ChatGPT and ChatGPT-Instruct, 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 respect to the input questions (Hwang et al., 2023), while our proposed ranking with LLMs can consider both input questions and their answer choices (Fig. 1). We further explore two key aspects: *(1) The alignment of two ranking 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 LLMs and orders sorted by semantic similarity scores, and the results are presented in Fig. 10. We find that the two ranking orders have minimal monotonous relations with means approximating 0 and low standard deviations. We also deep dive into cases with notable order variations to address (2). Appendix A.6 illustrates one such case in the `"Guns"` topic. We derive three observations. First, not all top-8 opinions by semantic similarity scores are helpful for predicting 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.

CoO versus CoT. Table 1 indicates that in predicting human opinions, Chain-of-Thought (CoT) (Wei et al., 2022) slightly hampers baseline DIO-top8 performance for ChatGPT and ChatGPT-Instruct. Conversely, Chain-of-Opinion reasoning (CoO) enhances overall performance. 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



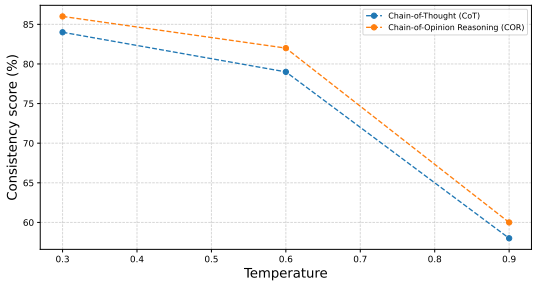


Figure 3: *Left:* Frequency distribution of accuracy over users by ChOiRe. *Right:* Consistency scores of the baseline DIO-top8 (ChatGPT) with CoO and CoT. The accurate consistency scores are in Appendix A.12.

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

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

the consistency score. The results are illustrated in Fig. 3. We observe that CoO brings slightly better consistent answers compared to CoT, especially when the temperature is high, verifying that CoO potentially enhances the reliability of the language models.

Dynamic Numbers of Opinions Analysis. Table 4 illustrates our extra analysis answering two research questions: *(1) How frequent can't LLMs answer the question?* and *(2) How do LLMs perform when more opinions than $K = 8$ are provided in ChOiRe?*. Our findings show that, firstly, with 8 opinions, GPT-4 exhibits the highest percentage of unanswered questions, while ChatGPT performs the best. 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 possibly harm the performance of models, our answer consistency strategy enables LLMs to achieve the best results across three different K values.

6.2 Error Analysis

FEA Misses Key Explicit Personae. Despite showing reasonably promising results in the task of 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 A.10. 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 A.6, we also observe they frequently include less relevant, or even irrelevant opinions to the ranked list such as in Appendix A.10-bottom. We attribute this to the challenge of this task, even for humans it might require substantial cognitive effort.

LLM 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 GPT-4. An illustrative example can be found in Appendix A.10-top-right. We posit that 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 reasonable and coherent text.

7 Conclusions

We propose ChOiRe, a four-step solution framework for individual opinion prediction via differentiating the utilization of user's explicit versus implicit personae. 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 sets up new strong SOTA results effectively with only limited inference calls. 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 explicit and implicit 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 personal information from explicit and implicit personae, which may be sensitive to some audiences and not be available. 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.

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.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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.

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.

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.

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.

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.

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

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.

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.

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.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1412–1421. The Association for Computational Linguistics.

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 L. 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. *Advances in Neural Information Processing Systems*, 35:27730–27744.

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.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

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.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. 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.

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.

Wordtune. 2022. Wordtune - chatgpt writing tool.

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 Appendix

A.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 (Wei 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: [...]
```

A.2 Prompt Templates for Implicit Feature Ranking

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

```
Given social behavior
question-answer pairs answered
by a user about his 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:
'{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: [...]
```

A.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.

A.3.1 W/o Persoba (Santurkar et al., 2023)

The W/o Persona prompt is provided below.

```
Question: {question}
Answer choices:
{choice}
Complete the answer by the following format:
Answer: A. or B. or C. or D. or E...
```

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

The DIO-top8 prompt 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, 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....
```

A.3.3 Chain-of-Thought (Wei 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:...
```

A.3.4 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:...
```

A.4 Top-3 Removed Explicit Personae Attributes

Table 5 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".

A.5 FEA Example with ChatGPT

Fig. 4 shows an FEA example with ChatGPT. We observe that by removing unnec-

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

Table 5: Top-3 explicit personae that got removed the most by the LLMs.

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 4: FEA example with ChatGPT.

essary explicit personae including "Age", "Citizenship", "Education", "Income", "Marital Status", "Race", "Frequency of religious attendance", ChatGPT predicts the opinion accurately, while without removing, a wrong prediction was made.

A.6 Example of High Disagreement between Rankings

Fig. 5 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 are helpful for predicting 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 a wrong prediction is made with the semantic similarity order.

A.7 Example of Inconsistent Answers Generated by CoT

Fig. 6 illustrates an example of the inconsistent answers generated by ChatGPT with Chain-of-Thought (Wei 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.

A.8 Example of Chain of Opinion Reasoning

Fig. 7 presents an example of the answer generated by ChatGPT using Chain of Opinion (ours) versus Chain of Thought (Wei et al., 2022) prompting methods.

Figure 5: Example of the high disagreement between orders by semantic similarity scores and LLM (ChatGPT).

A.9 Example of Answer Consistency with Dynamic Numbers of Opinions

Fig. 8 shows an example of the answer generated by GPT-4 using Chain of Opinion (ours) reasoning with different numbers of provided historical opinions.

A.10 Error Analysis Examples

Fig. 9 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.

A.11 Kendall's Tau Scores for Ranking Agreements

Fig. 10 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.

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 6: 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.

A.12 Consistency Scores

Table 6 presents the exact consistency scores for the Fig. 3. Besides CoO consistently outperforming CoT, we also observe that when the temperature is increased, the consistency score is decreased, which is intuitive.

A.13 Human Rating System

Our details of human rating instructions are provided in Table 7 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.

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

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

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 7: 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 8: 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"

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)

…
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.

Figure 9: Error analysis examples of ChOiRe with ChatGPT.

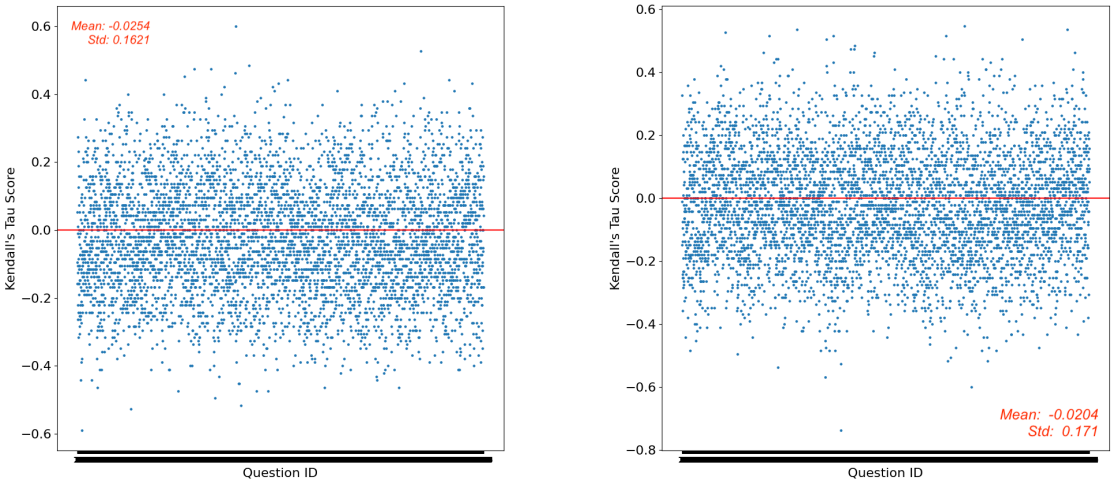


Figure 10: *Left:* Ranking agreements between ChatGPT top-K and semantic similarity top-K. *Right:* Between ChatGPT-Instruct top-K and semantic similarity top-K. One example that has a high disagreement score is shown in Appendix A.6.