A Graph per Persona: Reasoning about Subjective Natural Language Descriptions

Anonymous ACL submission

Abstract

Reasoning about subjective natural language descriptions such as opinions and preferences is a challenging topic which largely hasn't been solved to date. In particular, the state-of-the-art large language models (LLMs) perform disappointing in this task, show strong biases, and do not meet the interpretability requirements we often have in this kind of applications. We propose a novel approach for reasoning about subjective knowledge which integrates potential, implicit meanings and explicitly models the relational nature of the information. We apply supervised graph learning, offer explanations for the model's reasoning, and show that our model performs well across all 15 topics of OpinionQA, outperforming several prominent LLMs. Our detailed analysis further shows its unique advantages and the complementary nature it offers in comparison to LLMs.

1 Introduction

Subjective knowledge such as personal opinions and preferences represents a considerable challenge for automated reasoning. In fact, on the recently proposed OpinionQA datasets (Santurkar et al., 2023), even the state-of-the-art large language models (LLMs) reach surprisingly low scores and reveal certain biases (Santurkar et al., 2023; Hwang et al., 2023). As LLMs are incorporated into applications aimed at assisting individuals in daily tasks and decision making (OpenAI, 2023; Google, 2022; Ye et al., 2024), it is imperative that they can personalize their outputs for individual users.

One of the inherent problems with reasoning with subjective knowledge is its implicit nature. Rather than explicitly specifying their preferences and opinions, users may express these opinions indirectly through continuous interactions. Other properties that affect opinions and preferences may be external to the discourse, such as the demographic information and cultural background

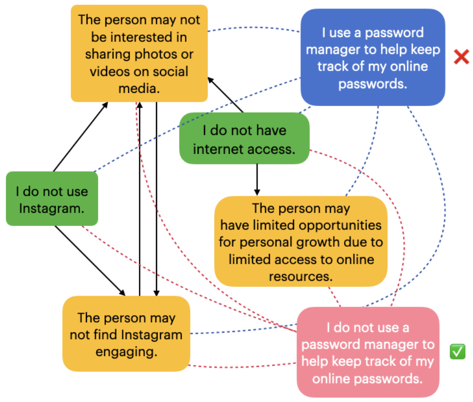

Figure 1: We model the relational nature of explicit and potential implicit opinions of an individual in a graph.

(Suriyakumar et al., 2023). Finally, we observe that various aspects of a problem are usually related, and the models often have to combine various pieces of information.

To test LLMs' ability to learn personal opinions, the OpinionQA dataset (Santurkar et al., 2023) presents models with the dialogue history containing a participant's responses to survey questions (e.g., `Do you use a password manager to help keep track of your online passwords?`), as well as their demographic information (e.g. Age: 50-64, Political affiliation: Republican). The model is then tasked with answering a set of multiple-choice questions pertaining to the opinions (e.g., `Yes/No`).

Current state-of-the-art LLMs still perform poorly on OpinionQA (Santurkar et al., 2023). In particular, models often ignore the survey history and over-rely on demographic information, which may lead to perpetuating societal biases (Hwang et al., 2023). Moreover, English LLMs struggle with questions from cross-national surveys (Durmus et al., 2023), given that they are trained on English web text coming primarily from users in the US. Current solutions focus on improving the reasoning by filtering the information that is available to the model when making a certain judgement (Hwang et al., 2023; Do et al., 2023), but there is still considerable room for improvement.

We propose an alternative approach to reasoning about subjective descriptions, inspired by traditional techniques modeling the relational nature of complex conceptual knowledge in semantic networks (Lehmann, 1992). Our framework, depicted in Figure 1, creates one opinion graph per individual, explicitly modeling relationships between their opinions on various topics (green). Due to the often intricate and implicit nature of opinions, we complete the graph with derived knowledge generated by a LLM (yellow). Finally, we add auxiliary nodes for the answer choices (blue, rose) and apply supervised graph learning to determine the opinions which are most relevant to the given question.

Our approach outperforms prominent LLMs across most of the 15 OpinionQA subsets. We ablate and evaluate our approach in detail. Most importantly, our analysis shows that our answers often complement those of the LLMs, which offers interesting future research potential. Finally, the graph neural network allows for extracting the attention flow over the graph nodes and hence naturally delivers an explanation for its reasoning. While the explanations are not perfect, they are useful for analyzing the reasoning steps and hint at future research questions.[1]

2 Related Work

Reasoning about Subjective Descriptions. Simpler forms of reasoning over subjective text have been studied in NLP for a long time in tasks such as sentiment prediction or user-item recommendation (Gao et al., 2023; He et al., 2017; Li et al., 2021). More complex tasks, predicting an opinion based on other opinions, have been considered recently with the study of personalized question answering over surveys (Santurkar et al., 2023; Durmus et al., 2023). Overall, LLMs have been shown to be underperforming (Santurkar et al., 2023; Ziems et al., 2023). Among their many findings, we point out the importance of curated personal opinions for personalized prediction (Hwang et al., 2023; Do et al., 2023). Understanding the model's ability to reason human opinions is crucial to ensure safer alignment with a user's ethical principles, moral beliefs, and cultural-specific values. We build upon the previous works by focusing on opinion data and employing graph learning to select opinions relevant to the task at hand.

Importance of Implicit Information. Most popular reasoning benchmarks focus on reasoning on objective knowledge. Additional factual context has been shown to improve LM reasoning in these setups (e.g., Akyürek et al., 2024). In the subjective context, we draw inspiration from the early work of Hobbs et al. (1988), who showed that explicit representations of meaning can help text understanding. More recently, Hoyle et al. (2022) showed importance of having explicit representations of implicit content with LLMs. We adopt this finding into our graph-based reasoning framework, which is an alternative to the popular chain-of-though reasoning paradigm (Wei et al., 2023; Yao et al., 2023; Besta et al., 2024), in which the LLM is reasoning in natural language. These methods often overly rely on demographic information when reasoning over human opinions, even in the presence of related opinions (Hwang et al., 2023).

Relational Reasoning. *"Relational reasoning, or the ability to consider relationships between multiple mental representations, is directly linked to the capacity to think logically and solve problems in novel situations"* in humans (Cattell, 1971; Crone et al., 2009). Motivated by this, graphs have been employed in NLP models to represent knowledge, primarily for reasoning about objective information (Jung et al., 2020; Xu et al., 2019; Das et al., 2021). To simulate step-by-step reasoning, Jung et al. (2020) and Das et al. (2021) particularly integrate reasoning paths in the models. We use the graph-based reasoning model from Jung et al. (2020).

3 Our Approach

Overview, Figure 2. Given a user's answers to previous opinion questions, our goal is to predict the answer to a multiple-choice question about an unstated opinion. We exploit the relational nature (entailment information) of personal opinions and create a graph for each person, containing their known opinions as nodes and, additionally, potential implicit meanings and relations between them; Sec 3.1. We encode the graph using graph embeddings. Specifically, we consider a supervised learning problem where the graph learner is biased to find paths leading from the nodes most relevant to the given question to possible answer nodes; the latter are added to the graph as auxiliary nodes;

[1] We will make the code available upon publication.

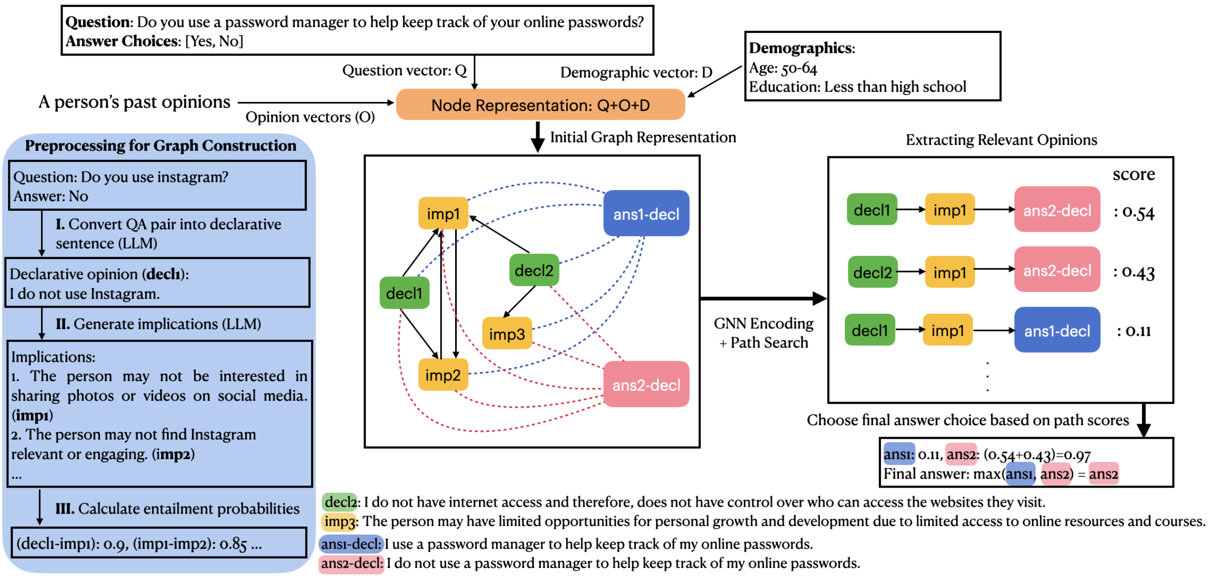

Figure 2: Overview of our approach: left: graph construction; middle: opinion graph; top: initial node embedding; right: extraction of reasoning paths, i.e., the relevant opinions, and answer calculation.

Sec 3.2. Lastly, we extract the highest-ranked paths to predict an answer; Sec 3.3.

Notation. We consider a given set of multiple-choice questions answered by a specific person: $\{(q_i, a_i, C_i)\}$ containing questions q_i, corresponding answer choices C_i, and the chosen answers $a_i \in C_i$. The question answering task is similarly given as a tuple (q, a, C) not part of the above set.

3.1 A Graph per Persona

I Given Opinions. We follow Hwang et al. (2023) and use the Wizard-Vicuna-30B model (Luo et al., 2023; Chiang et al., 2023) to convert each question-answer pair into a declarative sentence (e.g., `I do not use a password manager to help keep track of my online passwords.`). We obtain a set $\mathcal{O} = \{(q_i, a_i)\}$ representing the answers of a given survey participant and a set $\mathcal{T} = \{(q, c) \mid c \in C\}$ representing the task.

II Generating Implications. We use Wizard-Vicuna-30B to generate implications from the explicitly given opinions (see Appendix D for the prompt). For example, from the given statement: "`I do not use Instagram`", we can infer that the person may not be interested in sharing photos or videos on social media.

Since we observed some of the generated implications to be irrelevant in the context of the given opinion (see examples in Appendix E), we filter them as follows. We calculate the cosine similarity between the given opinion and each implication, and implications with a cosine similarity below a pre-defined threshold t_{sim} are discarded (we used $t_{sim} = 0.8$, based on preliminary experiments).

III Graph Construction. We construct a *multi-relational graph* $\mathcal{G} = (\mathcal{V}, \mathcal{R}, \mathcal{E})$ per person. The opinions $\mathcal{O}$ and implications $\mathcal{I}$ represent this set $\mathcal{V}$ of belief nodes, and we add the task encoding, i.e., $\mathcal{V} = \mathcal{O} \cup \mathcal{I} \cup \mathcal{T}$. For brevity, we often call all in $\mathcal{O} \cup \mathcal{I}$ *opinions*, although the implications are only potential derivations.

To capture an entailment relationship between opinions, we decide to represent the opinions as a graph structure. Since we generate multiple implications for each opinion, the graph should be dense by design. However, we are still missing more detailed knowledge about the exact nature of the connections (i.e., about the type or strength of the individual relations between two nodes). Specifically, we consider the set $\mathcal{R}$ of relation types to contain one type for opinion-opinion, opinion-implication, implication-opinion, and implication-implication edges, respectively, and define the set $\mathcal{E}$ of edges to contain all corresponding tuples $(v_i, v_j, r) \in \mathcal{V} \times \mathcal{V} \times \mathcal{R}$. That is, we have two edges between each pair of nodes, one in each direction; for uni-directional relationships, we add the corresponding two tuples.

Nevertheless, the implications are considered to be consequences of the opinions, and we assume additional such *entailment* relations to hold between other beliefs of the person. To model this information explicitly, we consider $\mathcal{R}$ to contain an additional entailment relation type. We compute these entailment edges using an LM, as described next.

We use a state-of-the-art model for natural language inference (NLI), T5-base (Raffel et al., 2020) to predict the probability p_{ij} for the graph edges to represent entailment.[2] We also use these predictions to filter out noise in terms of relationships, in that we consider the predicted entailment score to tell us about relatedness between our additional implications (and opinions) and filter out all edges below a pre-selected threshold t_{entail} (in our experiments, we chose $t_{entail} = 0.1$ based on manual observation). The final graph is thus no longer fully connected, but still dense enough for the model to broadly explore the space.

We fine-tuned the NLI model using the implications generated previously, since the model may lack prior knowledge about the specific domain under consideration (e.g., this was the case for the data we experimented with). More specifically, we consider each pair of opinion and corresponding generated implication as a positive example, and construct a single negative example for each positive example by pairing an opinion with a randomly chosen implication that was generated for another answer choice for the same question.

3.2 Reasoning over the Graph

Initial Graph Representation. For embedding the graph nodes, we apply a sentence embedding $\mathcal{M}_S : \mathcal{V} \rightarrow \mathbb{R}^{d_S}$ (we used Sentence Transformer[3]); a unique identifier for opinion nodes op : $\mathcal{V} \rightarrow \in \mathbb{R}^d$, which maps implications to the identifier of the opinion they were generated for; and a (binary) node type identifier typ : $\mathcal{V} \rightarrow \mathbb{R}^d$, which distinguishes opinion and implication nodes. We create an embedding as follows, for each $v_i \in \mathcal{V}$:

$$h_i^0 = W_v[\mathcal{M}_S(v_i)||\text{op}(v_i)||\text{typ}(v_i)],$$

where W_v represents a linear transformation.

The edge representations unify all relationships we have between a given pair of nodes v_i and v_j as follows: $e_{ij}^0 = W_e[e'_{ij}]$, W_e is a linear transformation, and e'_{ij} a one-hot vector with one flag per $r \in \mathcal{R}$. That flag is set to 1 if $(v_i, v_j, r) \in \mathcal{E}$; for the entailment relation, we set it to 1 if $p_{ij} > 0.5$, according to the predicted entailment probability.

[2]We also experimented with Flan-T5. T5 and Flan-T5 turned out to have similar performance in understanding the entailment relationship between subjective opinions.

[3]`BAAI/bge-base-en-v1.5`

Graph Learning using Graph Attention Flows. The goal in graph representation learning is to compute node representations h_j^t iteratively, for each layer t, by aggregating the embeddings h_i^{t-1} of the incoming neighbor nodes $v_i \in \overrightarrow{\mathcal{N}}_j$, i.e., $(v_i, v_j) \in \mathcal{E}$. The graph attention network (GAT) (Veličković et al., 2017) specifically applies attention to weigh the neighbors,[4] and there are versions taking relation types into account (Salehi and Davulcu, 2019). To emphasize the flow of information over the graphs, we follow works which compute the training loss by focusing on the attention values (Jung et al., 2020; Xu et al., 2019). The goal is to obtain attention values $\tilde{a}_i^t$ as a representation of the importance the answer choices have in the context of the opinion nodes. At each layer t (for readability we drop many superscripts $\cdot^t$):

- We first compute node embeddings using GAT:

$$\mathbf{h}_j^{t+1} = \sigma\left(\sum_{i\in\overrightarrow{\mathcal{N}_j}} a_{ij}\mathbf{W}_k(\mathbf{h}_j^t + \mathbf{e}_{ij}^t)\right)$$

$$a_{ij} = \text{softmax}_{i\in\overrightarrow{\mathcal{N}_j}}(e_{ij}^{t+1})$$

$$e_{ij}^{t+1} = \sigma((\mathbf{W}_n(\mathbf{h}_j^t + \mathbf{e}_{ij}^t)) \cdot (\mathbf{W}_m\mathbf{h}_i^t)^\intercal)$$

  where σ denotes leaky-ReLU and, for simplicity, j and v_j are used interchangeably.

- To bias the computation towards the question answering task under consideration, we incorporate a representation $\mathbf{q}$ of the target question, a sentence embedding acquired by Sentence Transformer. In case want to consider demographic features, we proceed similarly to obtain an embedding $\mathbf{d}$:

$$\hat{\mathbf{h}}_j^{t+1} = \mathbf{h}_j^{t+1} + \mathbf{W}_q\mathbf{q}(+\mathbf{W}_d\mathbf{d}).$$

- Instead of directly taking GAT's attention values as node importance scores, Jung et al. (2020) normalize them in the context of their neighbors and incorporate the values from previous steps. Note that initial scores $\tilde{a}_i^0$ then have to be given, we compute:

$$\tilde{a}_i^0 = h_i^0 \cdot (\mathbf{W}_q\mathbf{q} + \mathbf{W}_d\mathbf{d})$$

  To obtain normalized attention values $\tilde{a}_{ij}^{t+1}$ for each neighbor v_i, we weigh the edge from v_i to

[4]Observe that this can be seen as transformer architecture with a strong structural prior, in that attention for node pairs that are not connected by an edge are always 0.

Model	BERT	LLaMA-7b	Vicuna-13b	GPT-3.5	GPT-3	ChOiRe-ChatGPT	Mistral-7B	GOO
No Persona	-	0.33	0.36	0.37	0.43	-	-	-
op_{top8}	0.49	0.36	0.42	0.50	0.52	-	0.52	**0.55**
op_{top8}+demo	0.49	0.37	0.43	0.51	0.54	0.51	0.53	**0.55**

Table 1: Overall QA accuracy averaged over all OpinionQA datasets. No Persona: the LLMs run without any personalization; op_{top8}: given the 8 opinions most similar to the question (best for LLMs by Hwang et al. (2023)), for our model we use all; +demo: given demographics in addition.

v_j in the context of its outgoing neighbors $\overleftarrow{\mathcal{N}_i}$ and compute that impact γ_{ij} similar as above:

$$\tilde{a}_{ij}^{t+1} = \gamma_{ij}^{t+1}\tilde{a}_i^t$$
$$\gamma_{ij}^{t+1} = \text{softmax}_{j\in\overleftarrow{\mathcal{N}_i}}(\hat{e}_{ij}^{t+1})$$
$$\hat{e}_{ij}^{t+1} = \sigma((\mathbf{W}_{n'}(\hat{\mathbf{h}}_i^{t+1} + \mathbf{e}_{ij}^{t+1}))\cdot(\mathbf{W}_{m'}\hat{\mathbf{h}}_j^{t+1})^\intercal)$$

Note that here the $t+1$-step's node embedding impacts the node score. We obtain the final value by aggregating the incoming edges' values. Thus, a high score for the target node means it has a large influence onto its neighbors.

$$\tilde{a}_j^t = \sum_{i\in\overrightarrow{\mathcal{N}_j}} \tilde{a}_{ij}^t$$

Training Objective. We apply supervised learning as proposed by Jung et al. (2020); Xu et al. (2019), by focusing on the attention scores computed for the target answer node v_{target} across all layers $t \in T$. Note that this is because our data does not contain ground truth about which opinions are relevant to the task.

$$\mathcal{L} = \sum_{t=1}^{T} -\log \tilde{a}_{target}^t$$

3.3 Extracting Relevant Opinions

To determine the answer, we extract paths in the graphs with highest attention scores up to a depth T, considering each to contain opinions most relevant to the task; we chose $T = 3$. We collect these paths using a beam search, starting at $t = 0$ and consider the k nodes v_i with highest values $\tilde{a}_i^t$ and iteratively select the k neighbors with highest $\tilde{a}_i^{t+1}$ for each of them. We stop at $t = T$, drop all paths that do not end in an answer node, and score each remaining path P as follows:

$$s_P = \sqrt[|P|]{\prod_{t=0}^{|P|} \tilde{a}_{P(t)}^t},$$

where $|P|$ denotes the length of P, and $P(t)$ the index of the t-th node in P.

Then we obtain a score s_c per answer choice c, by aggregating the top-k scores of the paths $P \in \mathcal{P}_c^{\text{top-}k}$ leading to that answer; we used $k = 5$. Lastly, we select the highest one as the final answer.

$$\text{Ans}_c = \sum_{P\in\mathcal{P}_c^{\text{top-}k}} s_P$$
$$\text{Ans}_{\text{final}} = \max(\{\text{Ans}_c\})$$

We chose this prediction mechanism based on the top-k paths to include alternative sets (i.e., paths) of opinions into the prediction; we will also focus on the opinions in all top-k paths in our evaluation.

4 Evaluation

Settings. To test the model's personalization and reasoning ability, we use subsets of the 15 OpinionQA datasets (Santurkar et al., 2023) and train and test the models in a question-answering (QA) setup. In terms of baselines we consider BERT (Devlin et al., 2018), Mistral-7B (Jiang et al., 2023), text-davinci-003 (GPT-3), gpt-3.5-turbo (GPT-3.5), and ChOiRe (Do et al., 2023). The LLMs are used in a zero-shot setting. We use accuracy as primary performance metric. More details are given in Appendix A.

Overall Performance, Tables 1, 2, 3. At first glance, our models compete well with the LLMs. In particular, they show consistently good/best performance with and without demographic information. Among the LLMs this is only the case for GPT-3. We posit that the GPT3+ models trained on considerably larger datasets might have a better understanding of opinions.

We observe notable differences especially on *Guns*, *Biomedical-food*, and *Misinformation*. Comparing our models with and without implications, we observe that including implications significantly improves performance on most topics, particularly on *Sexual Harassment*, *Misinformation*, and *Global Attitudes*. Similarly, the entailment information further shows rather consistent performance improvements. Since the main table considers subsets

	Guns	Auto-mation	Gender	Sexual harass.	Biomed. food	Leadership	2050 US	Trust-Science
(L)LM BERT op$_{top8}$	62.5	47.5	54.1	37.7	54.3	52.3	42.9	57.3
Mistral op$_{top8}$	57.1	48.2	**56.1**	43.8	56.4	55.1	47.1	56.4
GOO op	$61.1_{1.2}$	$50.0_{0.5}$	$52.3_{1.1}$	$44.7_{1.9}$	$59.9_{1.9}$	$57.2_{0.8}$	$46.5_{0.9}$	$\mathbf{59.0}_{0.9}$
+imp	$\mathbf{62.1}_{1.2}$	$50.9_{0.4}$	$52.7_{0.3}$	$46.8_{0.5}$	$59.0_{1.1}$	$56.5_{1.3}$	$47.8_{0.2}$	$58.0_{1.0}$
+imp+entail	$60.8_{1.0}$	$\mathbf{53.5}_{1.1}$	$54.4_{0.9}$	$\mathbf{47.5}_{1.1}$	$\mathbf{61.3}_{1.4}$	$\mathbf{57.5}_{0.6}$	$\mathbf{49.8}_{0.5}$	$58.8_{0.7}$
LM op$_{top8}$+demo								
BERT	57.5	48.6	54.3	40.2	55.5	52.8	43.0	57.4
GPT-3.5	57.6	48.1	54.7	**47.9**	54.0	52.8	43.9	57.0
GPT-3	62.5	47.8	**57.0**	47.4	60.3	**59.1**	45.7	59.1
Mistral	57.0	51.0	55.7	45.6	55.3	57.2	49.1	58.0
ChOiRe-ChatGPT	57.1	49.2	59.2	39.9	54.7	52.2	49.5	56.4
GOO op+demo	$61.5_{1.4}$	$52.3_{0.9}$	$53.5_{1.1}$	$45.0_{0.2}$	$58.9_{1.8}$	$56.0_{0.8}$	$47.7_{1.5}$	$59.3_{0.3}$
+imp	$\mathbf{63.0}_{0.9}$	$52.0_{1.6}$	$54.4_{1.1}$	$46.7_{0.4}$	$\mathbf{61.2}_{0.3}$	$58.3_{1.6}$	$\mathbf{49.7}_{1.4}$	$\mathbf{60.0}_{0.4}$

	Race	Misinfor-mation	Privacy	Family	Economic Inequal.	Global Attitudes	Political Views	**Avg.**
(L)LM BERT op$_{top8}$	42.6	53.2	51.2	53.7	45.9	41.4	41.4	49.2
Mistral op$_{top8}$	49.1	48.9	**53.5**	55.5	51.2	49.5	47.7	51.7
GOO op	$51.6_{1.4}$	$54.7_{1.0}$	$50.3_{0.9}$	$55.5_{0.5}$	$53.0_{0.2}$	$48.8_{1.9}$	$55.0_{1.5}$	53.3
+imp	$\mathbf{51.8}_{0.8}$	$\mathbf{56.6}_{1.6}$	$50.4_{1.1}$	$56.3_{2.4}$	$52.8_{1.7}$	$\mathbf{53.3}_{1.6}$	$55.1_{0.3}$	54.0
+imp+entail	$51.2_{0.4}$	$56.0_{1.5}$	$52.3_{0.6}$	$\mathbf{57.3}_{0.6}$	$\mathbf{55.2}_{1.2}$	$52.0_{3.3}$	$\mathbf{55.3}_{0.6}$	**54.9**
(L)LM op$_{top8}$+demo								
BERT	46.2	52.0	47.8	51.8	46.0	42.5	43.7	49.3
GPT-3.5	50.1	48.0	51.0	54.9	49.5	47.2	48.5	51.0
GPT-3	51.0	54.5	51.1	57.0	**55.3**	48.2	51.6	53.9
Mistral	50.3	49.8	**53.9**	56.3	52.7	48.3	51.8	52.8
ChOiRe-ChatGPT	42.8	46.4	54.3	**60.0**	52.3	44.7	51.0	51.3
GOO op+demo.	$\mathbf{52.2}_{1.8}$	$54.4_{0.4}$	$50.0_{1.5}$	$52.6_{2.3}$	$51.6_{1.7}$	$\mathbf{52.8}_{0.3}$	$54.3_{1.1}$	53.5
+imp	$\mathbf{52.2}_{1.4}$	$\mathbf{56.9}_{0.7}$	$50.7_{1.0}$	$57.4_{0.7}$	$53.7_{0.5}$	$51.0_{1.0}$	$\mathbf{55.4}_{1.0}$	**54.8**

Table 2: Overall QA accuracy, top parts are without demographic information. Best in **boldface**, we color all those where the average is within the std. of the best, highlighting both the consistent performance across our models and the considerable differences to LLMs.

Model	Guns	Auto	Privacy
GOO op	$61.0_{0.5}$	$55.9_{0.7}$	$54.7_{0.3}$
+imp	$62.4_{0.5}$	$57.0_{0.3}$	$55.5_{0.1}$
+imp+entail	$\mathbf{62.5}_{0.7}$	$\mathbf{57.7}_{0.2}$	$\mathbf{56.4}_{0.4}$

Table 3: Scaling up the number of individuals.

of the data as they were used in previous works (Hwang et al., 2023; Do et al., 2023), we check what happens if we increase the number of survey participants whose answers we consider to 500, see Table 3. Interestingly, the positive impact of our proposed architecture gets more clear. Note that, in the setting with demographics, we do not consider the entailment version of our model since the entailment probabilities are computed for the original textual nodes but the demographic information, incorporated at each node, will likely change the nature of this relation.

Reasoning Examples, Figure 3. We start the analysis by showing an example, which also demonstrates the challenging nature of the problem. The figure shows the top-5 paths found leading to a correct answer prediction in **GOO**. Overall, we see that the fully-connected nature of the graph makes it possible to derive the answer directly based on a few relevant opinions (i.e., the paths are rather short). While these selected opinions seem all rather similar at first glance, observe that especially the derived, potential implicit meanings `The person may ...` point out interesting, often rather subtle aspects (e.g., possible political opinions, values more generally, or consequences on future plans). A more detailed error analysis is presented later, other examples are in the appendix.

Analyzing Predicted Relevant Opinions, Table 4. To give an impression of the nature of the explanations, we present statistics about the node types

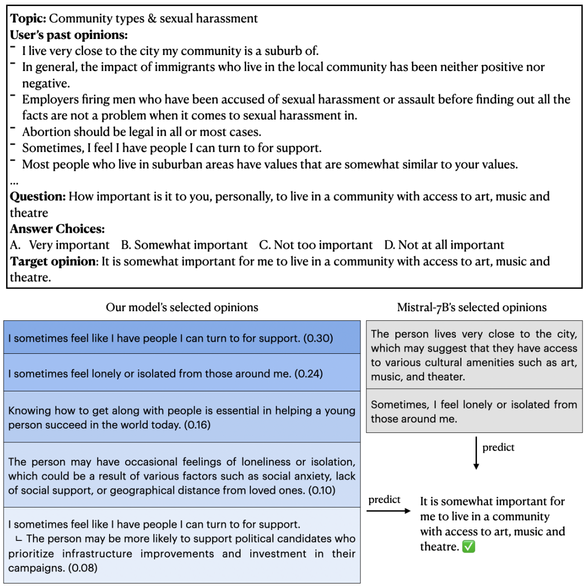

Figure 3: Example of most relevant opinions according to **GOO** op+imp+entail (path relevance scores) and Mistral-7B op_{top8}. "∟" denotes a next path node.

Model	# decl	# imp
Mistral-7B	2.7	-
GOO op+imp	2.7	2.1
+entail	2.6	2.2
+demo	2.8	1.9

Table 4: Average number of unique declarative opinions and implications in top-5 paths.

in the best paths, which reveal that they equally rely on explicit and implicit knowledge. The overall number of about three relevant opinions on average, plus two derived ones, seems reasonable. Observe that, when explicitly asked to explain its reasoning Mistral op gives a similar number of opinions. Table 9 in the appendix further shows the similarity in predicted relevant opinions in terms of overlap between different models and model variants on the correct predicted paths where both models agree. We see that the overlap between models can be rather low, which shows the need for making this information explicit and thus verifiable. These numbers also highlight that, in our model, adding entailment information can have more impact on the explanations than adding demographics. This underlines the power of this kind of implicit semantic and relational knowledge.

Moreover, we conducted a human evaluation comparing the outputs generated by our op+imp+entail model and Mistral-7B op through Amazon MTurk. We randomly selected 30 examples, two per topic, and each example was evaluated by three annotators. Annotators were asked to determine whether the target opinion could be inferred from a set of opinions chosen by our model (yes/no), along with a brief explanation. Based on the latter, we manually filtered out 13% noise. Overall 83% of our examples were deemed reasonable. Mistral-7B achieved a rating of 87%. However, note that the LLM was given the top-8 most similar opinions to the target question. Thus, finding relevant ones among those is much easier, and the scores are not directly comparable.

In what follows, we analyze the predicted answers in detail and show that both **GOO** and LLMs have unique advantages. Thus our work presents a promising, novel method to complement LLMs.

Comparing Individual Predictions, Table 5, Appendix B. We compute the agreement in correct and incorrect predictions between Mistral-7B and **GOO**. The numbers on a per-topic basis show that the trend is rather consistent and well reflected in the corresponding averages, 34/18/21/27%. This shows that the models may complement each other: When we combine the three cases where either of the models provides the correct answer, we can significantly improve the individual models' performance and obtain 73% accuracy.

We further show the agreement rates between the model variants, (e.g., **GOO** with and without entailment information) in Appendix B. Overall, we see that the agreement in both correct and incorrect predictions is considerably higher for variants of the same model than for different model families, both are around 40-50% across topics. First, this can be considered as verification that **GOO** is reasoning consistently in that adding information does not completely change the nature of the predictions. Interestingly, this is even the case where we compare the versions with(out) demographics for our model, but also for Mistral-7B in Table 7. Hence this also shows that combining different reasoning approaches (or model families) can be a promising direction to explore in the future.

Comparing Predictions on the Level of Individual Persons, Figure 4. The figure illustrates the distribution of how the model performs on a per-person basis, compared to Mistral-7B. We selected three topics where our model performs better than/similarly to/worse than Mistral-7B. The distributions from our model are generally less skewed, meaning that it shows more equal performance across individuals. In Mistral-7B, we

	Both	LLM	GOO	Both-X
Guns	0.39	0.18	0.21	0.21
Automation	0.33	0.15	0.21	0.31
Gender	0.38	0.18	0.16	0.27
Sexual harass.	0.25	0.18	0.23	0.33
Biomed. food	0.43	0.14	0.19	0.25
Leadership	0.35	0.20	0.23	0.22
2050 US	0.29	0.19	0.22	0.31
Trust-Science	0.40	0.17	0.20	0.23
Race	0.32	0.17	0.19	0.32
Misinfo.	0.31	0.18	0.27	0.24
Privacy	0.35	0.19	0.17	0.29
Family	0.37	0.19	0.21	0.23
Econ. Inequal.	0.33	0.18	0.22	0.27
Global Attitudes	0.33	0.16	0.23	0.28
Politics	0.31	0.17	0.25	0.27

Table 5: Agreement in predictions: both correct, only Mistral-7B $\text{op}_{\text{top-8}}$, only **GOO** op+imp+entail, both inc.

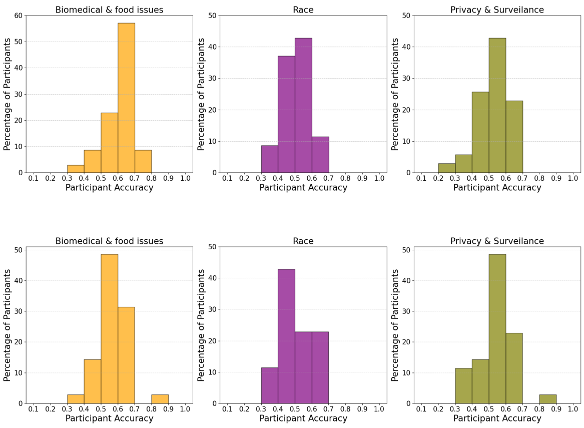

Figure 4: Accuracy-per-person distributions for **GOO** op+imp+entail (top) and Mistral-7B op_{top8} (bottom).

Model	all	rep.	dem.	ind.
Mistral-7B$_{\text{op-top8}}$	0.65	0.56	0.64	0.61
Mistral-7B$_{\text{op-top8+demo}}$	0.68	0.57	0.67	0.60
GOO op	0.75	0.68	0.70	**0.73**
+demo	0.74	0.68	0.68	0.70
+imp	0.74	0.66	0.70	0.71
+imp+demo	**0.76**	**0.69**	**0.73**	0.71
+imp+entail	**0.76**	0.68	0.71	0.69

Table 6: Overlap between model's majority answers and data's majority answers. all: entire data, rep.: republicans, dem.: democrats, ind.: independent.

observe that while the model achieves very high performance for certain people (*Biomedical-Food* and *Privacy*), resulting in an overall performance increase, there are also more individuals for which it's performing worse than our model. This experiment gives a more detailed view of how our model, or maybe even supervised learning more generally, could complement LLMs, to mitigate biases due to the potentially highly biased pre-training data.

Comparing Majority Predictions across Demographic Groups, Table 6. Here, we zoom out from the level of individual persons and consider the majority prediction of groups (i.e., all people in the dataset, and for groups with different political affiliations). Specifically, we compare them to the majority prediction from **GOO** and the LLM for those groups. There are interesting trends. First of all, the numbers are overall considerably higher for **GOO**, which makes it seem that the supervised approach allows the model to capture commonalities for certain populations, while this seems not the case for the LLM. Moreover, **GOO** does similarly well on all groups, even though the data itself is slightly biased (# rep./dem./ind.: 774/1075/683). On the other hand, the LLM, also here, shows clear bias (towards dem. opinions), even when given extra demographics. Overall, incorporating demographic information seems to generally enhance the models' ability to capture majority opinions.

Common Errors, Appendix C. We manually checked wrong predictions and corresponding explanations, see examples in the appendix. Amongst others, we noticed that including demographic information can overly strengthen a particular node and wrongly influence the selection of subsequent path nodes. Overall, we observe that the inclusion of demographics needs more careful consideration and study in future work. Furthermore, the diverse and nuanced context our graphs provide occasionally lead the model to irrelevant conclusions.

5 Conclusions

We propose a novel approach for reasoning about subjective natural language descriptions. Our approach represents a person's opinions in a graph which also includes generated implications, explicitly modeling the relationships between various statements. Given a question about a previously unstated opinion, we apply supervised graph learning to find a reasoning path from the existing knowledge to one of the candidate answers. Our model outperforms several prominent language models across all 15 topics of OpinionQA, while also offering explanations for its predictions. Detailed analysis further shows our model's unique advantages and the complementary nature it offers, in comparison to LLMs. Altogether, our work proposes a promising research direction to address this challenging problem and opens up interesting future research.

Limitations

From a data perspective, our work showed that we need better methods to integrate demographic or other additionally given information (i.e., beyond opinions), which is left as a challenging question for future research. We further note that our work, similar to the related works on the topic, focuses on the somewhat restricted survey scenario, where all users are captured in terms of one set of descriptions. If the latter varied (e.g., by having free-form answers), our supervised learning problem would become considerably harder. Our analysis has also clearly demonstrated that the implicit knowledge added using an LLM is often sensible, and manual checks are critical. Moreover, our approach is somewhat complex in that we need to apply an LLM during training for obtaining the derived knowledge; it is very efficient for inference though. For the LLM comparison, we applied a single prompt format as it was used in related works due to limited resources; ideally, we would average across a range of prompt templates. Finally, we point out that today's research (ours but also the related works) is far from being applicable in practice which, in turn, shows the critical need for this kind of research.

Ethics Statement

Data The dataset used in our work, OpinionQA (Santurkar et al., 2023) is publicly available. The dataset includes subjective opinions from humans and may contain offensive content to some people.

Data Collection We use Amazon Mechanical Turk to evaluate the quality of the opinions selected by our model and Mistral-7B. To ensure the quality of evaluation, we required that workers were located in English-speaking countries (e.g. US, UK, Canada, Australia, and New Zealand), and had an acceptance rate of at least 98% on 1,000 prior HITs. We paid $0.20 for the evaluation task. The annotators were compensated with an average hourly wage of $13, which is comparable to the US minimum wage. We did not collect any personal information from annotators.

Models The large language models we used for the experiments are trained on a large-scale web corpus and some of them utilize human feedback. This may also bring some bias when predicting user answers. With LLMs, users can select information that adheres to their system of beliefs and to amplify potentially biased and unethical views. Such an echo chamber (Del Vicario et al., 2016) can eventually cause harm by reinforcing undesirable or polarized a user's views. Our model based on a graph neural network mitigates these biases by focusing on the entailment relationship between opinions.

References

Afra Feyza Akyürek, Ekin Akyürek, Leshem Choshen, Derry Wijaya, and Jacob Andreas. 2024. Deductive closure training of language models for coherence, accuracy, and updatability.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2024. Graph of thoughts: Solving elaborate problems with large language models.

Raymond B Cattell. 1971. Abilities: Their structure, growth, and action.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Eveline A Crone, Carter Wendelken, Linda Van Leijenhorst, Ryan D Honomichl, Kalina Christoff, and Silvia A Bunge. 2009. Neurocognitive development of relational reasoning. *Developmental science*, 12(1):55–66.

Rajarshi Das, Manzil Zaheer, Dung Thai, Ameya Godbole, Ethan Perez, Jay Yoon Lee, Lizhen Tan, Lazaros Polymenakos, and Andrew McCallum. 2021. Case-based reasoning for natural language queries over knowledge bases. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9594–9611, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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. *Scientific reports*, 6(1):37825.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. Cite arxiv:1810.04805Comment: 13 pages.

Xuan Long Do, Kenji Kawaguchi, Min-Yen Kan, and Nancy F. Chen. 2023. Choire: Characterizing and predicting human opinions with chain of opinion reasoning.

Esin Durmus, Karina Nyugen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2023. Towards measuring the representation of subjective global opinions in language models.

Yunfan Gao, Tao Sheng, Youlin Xiang, Yun Xiong, Haofen Wang, and Jiawei Zhang. 2023. Chat-rec: Towards interactive and explainable llms-augmented recommender system.

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

Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering.

Jerry R. Hobbs, Mark Stickel, Paul Martin, and Douglas Edwards. 1988. Interpretation as abduction. In *26th Annual Meeting of the Association for Computational Linguistics*, pages 95–103, Buffalo, New York, USA. Association for Computational Linguistics.

Alexander Miserlis Hoyle, Rupak Sarkar, Pranav Goel, and Philip Resnik. 2022. Are neural topic models broken? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5321–5344, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

EunJeong Hwang, Bodhisattwa Majumder, and Niket Tandon. 2023. Aligning language models to user opinions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5906–5919, Singapore. Association for Computational Linguistics.

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.

Jaehun Jung, Bokyung Son, and Sungwon Lyu. 2020. AttnIO: Knowledge Graph Exploration with In-and-Out Attention Flow for Knowledge-Grounded Dialogue. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3484–3497, Online. Association for Computational Linguistics.

Fritz Lehmann. 1992. Semantic networks. *Computers Mathematics with Applications*, 23(2):1–50.

Shuyang Li, Bodhisattwa Prasad Majumder, and Julian McAuley. 2021. Self-supervised bot play for conversational recommendation with justifications.

Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xiubo Geng, Wenxiang Hu, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. 2023. Wizardcoder: Empowering code large language models with evol-instruct.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

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

Amin Salehi and Hasan Davulcu. 2019. Graph attention auto-encoders.

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.

Vinith Menon Suriyakumar, Marzyeh Ghassemi, and Berk Ustun. 2023. When personalization harms performance: Reconsidering the use of group attributes in prediction. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 33209–33228. PMLR.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2017. Graph attention networks. *6th International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Xiaoran Xu, Songpeng Zu, Chengliang Gao, Yuan Zhang, and Wei Feng. 2019. Modeling attention flow on graphs.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models.

Yining Ye, Xin Cong, Shizuo Tian, Yujia Qin, Chong Liu, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2024. Rational decision-making agent with internalized utility judgment.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2023. Can large language models transform computational social science?

A More Details about Evaluation Settings

Dataset To test the model's personalization and reasoning ability, we use the OpinionQA dataset (Santurkar et al., 2023) and train and test the model under a question-answering (QA) setup. OpinionQA dataset contains 15 topics ranging from guns, global attitudes, and political views, and each topic contains an average of 100 questions and 5K users. Due to limited resources, we follow previous works (Hwang et al., 2023; Do et al., 2023) that use sampled data, in which the data includes 100 users per topic and each user has their past opinions up to 16 and 30 opinions to evaluate the model's personalization and reasoning capabilities. Then, we use 35 users per topic to test the model's abilities, ensuring the same test set used in Hwang et al. (2023), and the rest are used as training. The final dataset results in a total of 525 users and 45K QA pairs. In our setting, we treat political ideology information as a part of user demographics.

Baselines We compare our model performance with BERT (Devlin et al., 2018), Mistral-7B (Jiang et al., 2023), text-davinci-003 (GPT-3), gpt-3.5-turbo (GPT-3.5), and ChOiRe (Do et al., 2023). BERT is a transformer-based language model, which can be finetuned for a wide range of tasks, including question answering and natural language inference. In our task, input to the BERT model is: `[USER user id][DEMO]demographics[SEP][OPINION]topk opinions[SEP]question` and the model is trained to predict the user's answer for a given question. Mistral-7B (Jiang et al., 2023) is a large language model that improves generation quality and facilitates inference using grouped-query attention and sliding window attention. Mistral-7B performs on par with LLaMA2-13B and LLaMA-34B (Touvron et al., 2023), across diverse tasks, including reasoning. LLaMA1 and LLaMA2 are transformer-based language models that were trained on trillions of tokens from exclusively publicly available data. ChOiRe (Do et al., 2023) is an approach with a chain of opinion reasoning. They propose a 4-step framework that filters out irrelevant information in demographics or user opinions to answer an input question.

Metric For accuracy evaluation, we simply calculate the accuracy of the predicted answer choice to the gold answer choice in the dataset.

Hyperparameters We use 5 implications for each opinion. The number of GAT layers was set to 3. When selecting top-k paths, we set K to 5. The learning rate is set to 0.00005, the number of epochs is set to 30, and the batch size is set to 1 due to a varying number of opinions for each user. We used GPU A40 for all our experiments and our model took 2-3 hours. Our models ran three times with different seed numbers and we report the average of them with their standard deviations.

	Both	LLM1	LLM2	Both-X
Guns	0.50	0.07	0.07	0.36
Automation	0.41	0.07	0.10	0.42
Gender	0.48	0.08	0.08	0.36
Sexual harass.	0.33	0.11	0.12	0.44
Biomed. food	0.48	0.09	0.08	0.36
Leadership	0.49	0.06	0.09	0.36
2050 US	0.40	0.08	0.10	0.43
Trust-Science	0.50	0.06	0.07	0.36
Race	0.42	0.07	0.08	0.43
Misinfo.	0.41	0.08	0.09	0.42
Privacy	0.47	0.06	0.07	0.40
Family	0.48	0.08	0.09	0.36
Econ. Inequal.	0.43	0.09	0.10	0.39
Global Attitudes	0.40	0.09	0.08	0.43
Politics	0.38	0.10	0.14	0.38

Table 7: Agreement in individual predictions: both correct, only Mistral-7B $\text{op}_{\text{top-8}}$, Mistral-7B $\text{op}_{\text{top-8+demo}}$, both incorrect.

B Additional Results: Comparing Predictions

Table 7 and 8 show agreement rates in individual predictions among the same model variants (e.g Mistral-7B $\text{op}_{\text{top-8}}$, Mistral-7B $\text{op}_{\text{top-8+demo}}$)

C An example of common errors

Figure 5 shows a common error when incorporating demographics.

	Both	GOO1	GOO2	Both-X
Guns	0.54	0.07	0.08	0.31
Automation	0.51	0.03	0.03	0.43
Gender	0.48	0.06	0.07	0.39
Sexual harass.	0.38	0.10	0.10	0.41
Biomed. food	0.52	0.10	0.08	0.31
Leadership	0.51	0.07	0.09	0.33
2050 US	0.39	0.12	0.08	0.41
Trust-Science	0.59	0.00	0.01	0.40
Race	0.43	0.08	0.10	0.39
Misinfo.	0.51	0.07	0.05	0.37
Privacy	0.47	0.05	0.05	0.43
Family	0.51	0.08	0.10	0.32
Econ. Inequal.	0.37	0.18	0.11	0.34
Global Attitudes	0.47	0.09	0.07	0.37
Politics	0.46	0.10	0.09	0.35

Table 8: Agreement in individual predictions: both correct, only GOO op+imp, GOO op+imp+demo, both incorrect.

Model	Opinion Overlap
op+imp vs. Mistral-7B	0.18
op+imp+entail vs. Mistral-7B	0.12
op+imp vs. op+imp+demo	0.41
op+imp vs. op+imp+entail	0.26

Table 9: Opinion overlap between different model variants in the top-5 paths

D Prompt for generating implications

To generate implications for opinions, we use the following prompt:

```
USER: For a question: <question> with
the following answer choices: [<choice1>,
<choice2>, <choice3>], a person chose
<choice1> as the answer. What does this imply?
Generate implications in up to 5 sentences.
1. <implication1>
2. <implication2>
3. <implication3>
4. <implication4>
5. <implication5>
ASSISTANT:
```

E Examples of irrelevant implications

F Prediction Distribution on More Users

Figure 7 presents the distribution of how the model performs on 100 users. We observe a similar trend to the distributions with 35 users.

G Amazon MTurk for human evaluation

For human evaluation, we instruct annotators as follows:

> You will be given a survey question, a person's answer choice for the question, and their past opinions. Evaluate whether the selected opinions are reasonable to address the person's answer choice for a given question.

Next, we present Figure 8 to annotators. Annotators are asked to evaluate the quality of selected opinions with a short explanation of why. We conduct two rounds of evaluation (our model and Mistral-7B) to avoid annotators being biased by looking at the responses from another model variant.

Question: Still thinking ahead 30 years, which do you think is more likely to happen in the U.S.? The U.S. economy will be stronger/weaker

Choices:
The U.S. economy will be stronger
The U.S. economy will be weaker

Opinions:
The respondent believes that Social Security benefits should not be reduced in any way when thinking about the long-term future of Social Security.
Increasing spending for roads, bridges, and other infrastructure is a top priority for improving the quality of life for future generations according to the respondent.
...

Selected paths w/ opinions:
- Increasing spending for roads, bridges and other infrastructure should be a top priority for the federal government to improve the quality of life for future generations. (0.51)
- If I were deciding what the federal government should do to improve the quality of life for future generations, I would give reducing the national debt an important but not top priority. (0.21)
- Increasing spending for roads, bridges and other infrastructure should be a top priority for the federal government to improve the quality of life for future generations.
  ∟ Thinking about the long-term future of Social Security, I think social Security benefits should not be reduced in any way. (0.16)
Providing high-quality, affordable health care to all Americans should be a top priority for the federal government to improve the quality of life for future generations. (0.15)
...

Selected paths w/ opinions + demographics:
- The automation of jobs through new technology in the workplace has neither helped nor hurt overall. **(0.68)**
- The automation of jobs through new technology in the workplace has neither helped nor hurt overall.
  ∟ The person who chose "Major problem" may be more likely to be aware of the prevalence of sexual harassment and assault in the workplace and may be more likely to take steps to prevent it from happening (0.07)
The automation of jobs through new technology in the workplace has neither helped nor hurt overall.
  ∟ The person may be more likely to support the idea that employers should take a more active role in preventing and addressing sexual harassment and assault in the workplace (0.07)
...

User-answer (expected): Weaker
Model with opinions: Weaker ✓
Model with opinions+implications: Stronger ✗

Figure 5: An example of demographics affecting the model's start node. As observed in chosen paths with opinions+demographics, demographic information can excessively emphasize irrelevant details, causing subsequent nodes in the path to lose relevance with input question.

Question: Please think about what things will be like in 2050, about 30 years from now. Thinking about the future of the United States, would you say you are
Choice: Very optimistic
Converted Declarative opinion: I am very optimistic about the future of the United States in 2050.

Relevant Implications:
The person may be more likely to take actions that contribute to a positive future, such as supporting sustainable practices or participating in democratic processes.

The person may be more likely to seek out information and news that reinforces their positive outlook.
...

Irrelevant Implications:
The person may be more likely to engage in activities that promote positive thinking, such as meditation or mindfulness practices.

Figure 6: Example of irrelevant implication with respect to the given converted declarative opinion generated by Wizard-Vicuna-30B. We filter out such irrelevant implications.

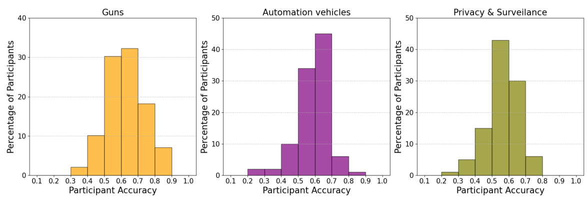

Figure 7: Accuracy-per-person distributions for **GOO** op+imp+entail on 100 people.

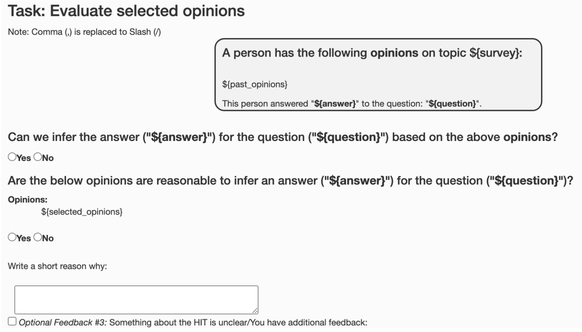

Figure 8: Amazon MTurk Screen for human evaluation to evaluate the quality of selected opinions.