Evaluating Large Language Models for Belief Inference: Mapping Belief Networks at Scale

Trisevgeni Papakonstantinou[♡] **Antonina Zhiteneva**[♢] **Ana Ma**[△]
Derek Powell[△] **Zachary Horne**[□]

[♡]University College London [♢]University of Oxford
[△]Arizona State University [□]University of Edinburgh
ucjutpa@ucl.ac.uk

Abstract

Beliefs are interconnected, influencing how people process and update what they think. To study the interconnectedness of beliefs at scale, we introduce a novel analytical pipeline leveraging a finetuned GPT-4o model to infer belief structures from large-scale social media data. We evaluate the model's performance by (1) comparing it to human annotated data (2) and its inferences to human-generated survey data. Our results show that a fine-tuned GPT-4o model can effectively recover belief structures, allowing for a level of scalability and efficiency that is impossible using traditional survey methods of data collection. This work demonstrates the potential for large language models to perform belief inference tasks and provides a framework for future research on the analysis of belief structures.

1 Introduction

Beliefs do not exist in isolation, they co-occur and cohere with each other, shaping how people process information and update their views (Brown, 2022; Brown and Kaiser, 2021; Brandt and Sleegers, 2021; Enders et al., 2024). Our understanding of the world is encoded with intricate belief structures – networks of individual beliefs and their interrelations, including points of coherence, contradiction, and overlap within broader systems of attitudes. Understanding these connections can be key to educational efforts in domains such as vaccine hesitancy and climate change (Powell et al., 2022; Schotsch and Powell, 2022).

Traditional approaches to studying belief structures rely on controlled surveys and Likert scales to measure the co-occurrence of beliefs, offering snapshots of belief structures for limited sets of domains (e.g., vaccine attitudes). In contrast, social media data from platforms like Reddit provide vast unstructured data where people express and debate their views, offering an opportunity to develop methods to infer beliefs directly from language. While some emerging work has leveraged large-scale social media data to study how people change their beliefs (Priniski and Horne, 2019; Priniski and Holyoak, 2020; Papakonstantinou and Horne, 2023; Priniski and Horne, 2018; Tan et al., 2016), these studies have not tackled the difficult methodological constraint of directly modeling belief structures from unstructured data. Analysing belief connections reveals the structure of belief systems, which is key to understanding and influencing attitudes. This structural insight enables more effective interventions, as shown in prior work on vaccine attitudes and belief polarisation (Powell et al., 2022; Horne et al., 2015; Cook and Lewandowsky, 2016)

We bridge this gap by developing and validating a framework to identify people's beliefs from online posts, using data from *ChangeMyView* (CMV), a Reddit forum where users debate and revise their beliefs. Unlike previous studies that rely on predefined belief measures or focus narrowly on single topics, our work centers on the core challenge of inferring individual beliefs and their co-occurence across users from naturalistic text. Moreover, while argumentation and persuasion have been extensively studied in the CMV dataset (Priniski and Holyoak, 2020; Priniski and Horne, 2019, 2018; Papakonstantinou and Horne, 2023; Tan et al., 2016), what has not been done —and what we address in this work— is the structured extraction of users' belief positions across multiple topics. We focus on the novel challenge of inferring what people believe and how those beliefs cohere in a scalable manner, rather than patterns of argumentation and belief change.

Efforts to align LLMs with human beliefs have shown potential for applications such as virtual surveys and behavioural modeling (Namikoshi et al., 2024). Yet, there is no established framework for evaluating a language model's ability to infer people's underlying beliefs in a way that allows us to

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 20787–20795
November 4-9, 2025

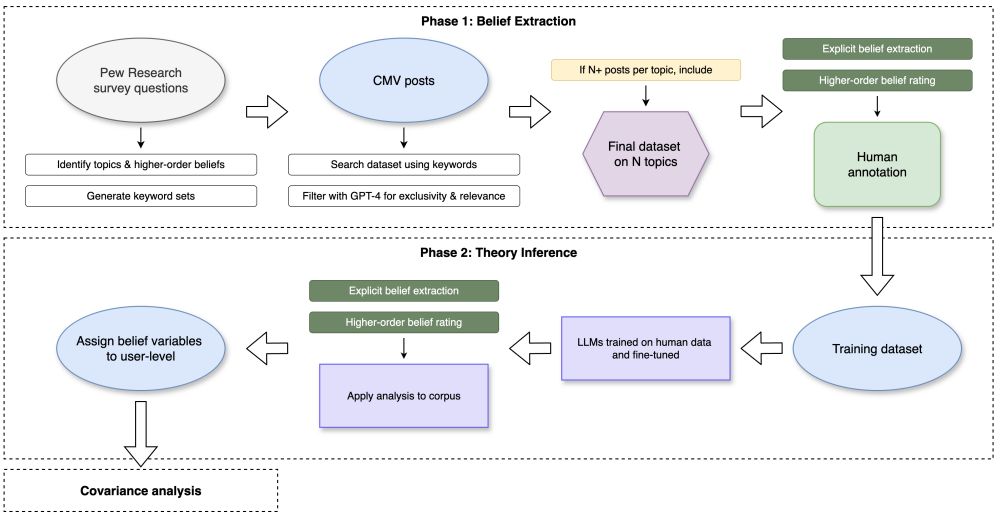

Figure 1: Detailed workflow of the data pre-processing steps for the belief extraction and theory inference process prior to the LLM annotation at scale.

recover their structure – how they co-occur and cohere. While previous research has explored some related dimensions in a variety of NLP tasks (e.g., consistency and uniformity), no existing studies have explicitly tested whether LLMs are able to infer structured belief representations across multiple topics in noisy naturalistic data.

To address this, we apply large language models (LLMs)—here, a fine-tuned GPT-4o model—in a robust analytical pipeline designed to recover social media users' belief networks from their Reddit posts. Figure 1 presents a description of the steps involved in this pipeline. We compared the performance of a baseline GPT-4o model, a fine-tuned GPT-4o model, and OpenAI o3-mini, finding that the fine-tuned version of GPT-4 performs as well as humans performing the same task. Specifically, we find that this pipeline, a fine-tuned version of GPT-4o accurately recovers users' beliefs. When we qualitatively compare the model's performance against a snapshot of human survey data, we show that it recovers the correlation between peoples' beliefs across domains even when making inferences on noisy and sparse data. This work advances both the study of belief networks and the use of LLMs in large-scale belief inference.

2 Methods

This study explores the feasibility and performance of an automated analytical pipeline using GPT-4o to accurately infer belief structures from specific subreddit posts, compared to ground truth derived from human annotations.

2.1 Data and Materials

ChangeMyView Dataset We collected data from the CMV subreddit, downloaded from Academic Torrents and ranging submissions from June 2005 to June 2024. The posts were scraped by pushshift and *u/RaiderBDev* (`stuck_in_the_matrix`). We constructed the final dataset through a two-step filtering process focused on a predefined set of 64 belief topics (e.g., abortion rights) derived from Pew Research surveys. Posts were initially retrieved using topic-specific keyword searches and subsequently refined using GPT-4o to remove noise, resulting in a curated dataset of 3,082 posts from 346 users. (See Appendix for full details.)

2.2 Analytical Pipeline

We propose and validate an analytical pipeline that takes raw text from the Reddit posts and infers the beliefs held by the user who authored the post, based on the statements made within it. We constrain the task by only allowing the beliefs inferred by the language model to fall within a predefined set of possibilities, with each post classified under a specific belief topic.

This process proceeds in two phases, reflecting a separation between belief extraction and belief inference, which are fundamentally different tasks with distinct output formats. In Phase 1, the task is to extract explicit belief statements from the posts. This involves producing standalone belief statements (e.g. "there is not a single logical reason why

gay marriage shouldn't happen"; see Appendix Table 2), which are then paired with corresponding ratings during the next phase (described below). Fine-tuning a model specifically for this extraction task ensures high fidelity in capturing belief content and yields a structured dataset of belief statements aligned with human annotations. In Phase 2, the task shifts to interpreting these extracted statements by mapping them onto the broader, pre-specified belief dimensions using a Likert scale (1–7), to mirror the task of human raters filling out survey questions.

We performed stratified random sampling to validate this pipeline with human annotations. First, we selected belief topics at random and then sampled relevant posts within that topic. For example, we randomly selected among topics ranging from gun control to abortion rights and then chose posts relevant to those topics.

Phase 1: Belief Extraction A sample of 200 posts derived from 20 belief topics were independently annotated by two human annotators. In this stage, the annotation required the human annotators to extract utterances within the post that the annotators agreed represented explicit belief statements. For example, the statement "I don't believe that there are any logical reasons why gay marriage shouldn't be allowed" was coded as an explicitly expressed belief, but the statement "when I try to understand the basis for the argument of the other side, I see people bashing it because of their personal beliefs and religious morals" was not coded as a belief (see Appendix Table 2 for an example of the input and output of this phase). We aimed to extract a list of standalone belief statements while preserving the original wording as much as possible. This approach ensured that each extracted belief could be interpreted independently of its surrounding context, facilitating downstream analysis and enabling the creation of a dataset suitable for tasks such as modeling belief attribution. By retaining the original phrasing wherever possible, we aimed to preserve the author's intent and minimise potential bias introduced through rewording. This process produced the "ground truth" dataset, where we determined the mapping of each post to a set of explicit belief statements based on the labeling of human annotators.

Phase 2: Theory Inference This phase involved mapping the extracted belief statements from Phase 1 to the pre-specified broader beliefs taken from a Pew research survey. Each set of belief statements corresponding to a post, as coded in Phase 1, was evaluated on whether (1) it agreed with a pre-specified broader belief, (2) agreed with the negation of the broader belief, and (3) neither agreed nor disagreed with it. These dimensions of agreement were evaluated independently. For example, the Pew research survey included the broader belief topic that the American economic system unfairly favors powerful interests. In this case, three annotators evaluated a post on the following three dimensions: "The economic system unfairly favors powerful interests," "The economic system does not unfairly favor powerful interests," and "The economic system doesn't have a clear positive or negative bias toward powerful interests." That is, three human annotators independently rated whether, based on the beliefs extracted in Phase 1, a poster endorsed each broader belief statement taken from the Pew research survey. This was done using a seven-point Likert scale to mirror the rating scales used by Pew. Disagreements were resolved through discussion between the human annotators, and when ratings fell within two points of each other, the mean of the three was used as the final likelihood label (for similar procedure for resolving disagreements, see Eagly and Revelle, 2022). This process produced a robust training and evaluation dataset with substantial agreement across annotators (Krippendorf's α = 0.69).

Fine-Tuning We finetuned GPT-4o to predict the likelihood of alignment with a broader belief for each pre-processed CMV post using the OpenAI finetuning API (see Appendix for details). The input to the model consisted of users' posts in the form of extracted belief statements, as described in Phase 1 (i.e., text data) along with their Phase 2 belief annotations in the format of a Likert scale point (an integer from 1-7).

3 Evaluation

We evaluated three models on the belief inference tasks described above: Baseline GPT-4o, a fine-tuned GPT-4o, and o3-mini. This evaluation was conducted on a dataset of 200 posts covering 20 distinct topics, with an 80/20 train-test split and employing 5-fold cross validation. Each model's performance was compared against our human-annotated ground truth dataset. We report ± 1 accuracy to capture near-miss errors in ordinal be-

lief strength predictions, Cohen's κ to assess inter-annotator agreement, Spearman's ρ to evaluate rank correlation across belief scores, and cosine similarity to assess vector-level structural similarity between belief representations.

Results As shown in Table 1, our fine-tuned GPT-4o model achieved the best performance across all metrics. Fine-tuning dramatically improved the model's ability to perform the belief inference task. Notably, the fine-tuned GPT-4o performs at a level comparable to human annotators.

Generalizing task performance to novel beliefs and users We further explored the robustness of the finetuned GPT-4o to the belief inference task by examining its ability to infer novel beliefs in domains it was not trained on. To examine this, we conducted a second evaluation under a cross-validation procedure defining train-test splits based on non-overlapping belief topics. We held out 4 of 20 beliefs for validation, ensuring the model was evaluated on entirely unseen thematic content. This setup assesses the models ability to generalize beyond the topical distribution of the training data, simulating a real-world deployment scenario where the model is asked to infer beliefs beyond those it has been explicitly trained on. The fine-tuned GPT-4o model maintained strong performance under this split (± 1 Accuracy = 86.67%, Cohen's κ = 0.85, Spearman's ρ = 0.88, Cosine Similarity = 0.96).

We finally examined user-level generalization as another important measure of robustness. We conducted a held-out users evaluation to ensure that reported performance reflects the model's ability to generalize to entirely unseen users, not just unseen posts or topics. Under this setup, the fine-tuned GPT-4o again demonstrated robust performance, though with some expected degradation: ± 1 Accuracy = 72.97%, Cohen's κ = 0.73, Spearman's ρ = 0.68, Cosine Similarity = 0.91.

4 Correlation Analysis Results

Comparison with Lab Sample One use case for belief extraction from naturalistic text are efforts to better understand the systematic connections among people's beliefs or attitudes (Brandt and Sleegers, 2021; Powell et al., 2023; Priniski and Horne, 2018). As a preliminary test of these applications, we applied our belief extraction pipeline to estimate the structure of belief correlations among CMV posters. To validate our results, we compared them against a secondary dataset of correlations computed within a survey of U.S. respondents (Ma and Powell, 2025).

Notably, there are substantial challenges to inferring relationships between people's beliefs based on their noisy and sparse expression of those beliefs in an online forum. Chiefly, most users post infrequently, and so do not post about the vast majority of their beliefs. We approach this problem by computing correlations among extracted belief scores in a pairwise fashion while ignoring missing data from posters who never posted about a topic.

We compared the correlations derived from the fine-tuned GPT-4o's ratings with those obtained from the survey (Figure 2). We calculated the Pearson correlation coefficient between the two correlation matrices was found to be $r = 0.56$, indicating a moderate level of agreement.

5 Conclusions

We find that the analytic pipeline described above allows a fine-tuned version of GPT-4o to achieve high reliability in inferring the beliefs people express, in turn allowing us to recover the co-occurence and coherence of these beliefs, and can do so even when its input is noisy and sparse social media data. These findings serve as an illustrative proof of concept, suggesting that our pipeline might be applied to explore belief structures in novel contexts. By leveraging scalable methods, our approach enables the analysis of beliefs in on-

	±1 Accuracy (SE)	Cohen's κ	Spearman's ρ	Cosine similarity
Baseline GPT-4o	53.12% (1.48)	0.35	0.43	0.85
Fine-tuned GPT-4o	**82.22% (2.82)**	**0.78**	**0.72**	**0.93**
o3-mini	65.56% (1.41)	0.40	0.42	0.81
Human	80.68% (1.83)	0.73	0.70	0.91

Table 1: Evaluation of models based on accuracy, reliability, correlation and similarity for belief inference tasks against human ground truth annotation. The SE for the fine-tuned GPT-4o represents variance between folds

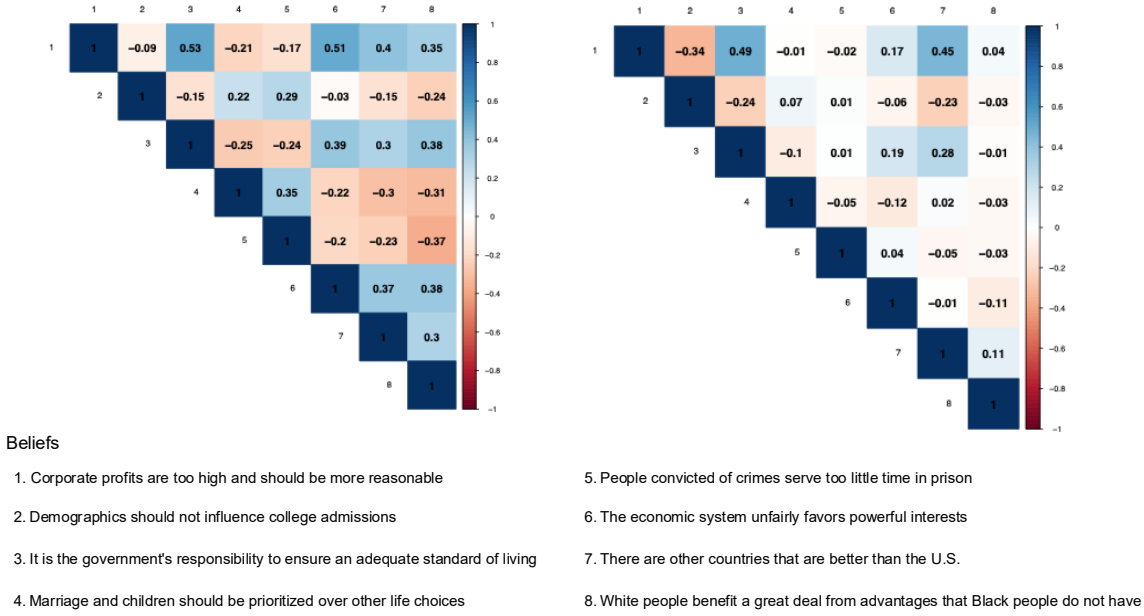

Figure 2: Heatmap of Spearman's correlation coefficients between beliefs in the survey data (left) and CMV data (right). The strength and direction of relationships are represented by the color gradient, ranging from strong negative (*dark red*, $r = -1$) to strong positive correlations (*dark blue*, $r = 1$)

line settings, offering a powerful tool for tracing belief networks from real-world text.

Limitations

Representativeness of Data Reddit users, and particularly those active on CMV, are not representative of the broader population. The platform skews heavily towards young, white, middle-class American men, meaning the beliefs expressed, and the co-occurrence patterns we observe, are likely shaped by a relatively homogenous set of worldviews. As such, care should be taken in generalising these results beyond this specific online context to more diverse populations with different cultural or socio-political backgrounds.

Sparsity of Belief Signals Most users only post about a narrow subset of the belief topics we study, meaning their full belief structure is only partially observable. As a result, many inferred belief relationships are necessarily incomplete.

Models Evaluated While the fine-tuned GPT-4 model performs on par with human annotators, our evaluation is limited only to OpenAI models. Different architectures, training data, or fine-tuning strategies may yield different results.

Inference Circularity While both the filtering and inference stages used GPT-4o, we used the base model for filtering and a fine-tuned model was evaluated and applied for the belief inference task. The fine-tuned GPT-4o performed dramatically better than the base model on belief inference, indicating that it learned beyond the filtering model's capabilities. This helps reduce concerns about circularity, since the belief inference model was not used to determine what data it would later see.

Qualitative Alignment to Human Data Our comparison between model-inferred belief correlations and those from the survey data was intended as a qualitative measure of alignment, based primarily on observation of the covariance matrices. However, we acknowledge that this approach does not explicitly account for important factors such as differences in sample sizes, missing data, or uncertainty in the belief estimates. These limitations constrain the strength of any conclusions drawn from this comparison. Nonetheless, the pipeline we've laid out suggests that this enterprise is both feasible and valuable: our results show that large language models can recover meaningful belief structure from unstructured text in a way that broadly aligns with population-level trends.

References

Mark J Brandt and Willem WA Sleegers. 2021. Evaluating belief system networks as a theory of political belief system dynamics. *Personality and Social Psychology Review*, 25(2):159–185.

Anna Brown. 2022. Deep partisan divide on whether greater acceptance of transgender people is good for society.

R Khari Brown and Angela Kaiser. 2021. Religious ideology, race, and health care policy attitudes. *Politics and Religion*, 14(4):764–786.

John Cook and Stephan Lewandowsky. 2016. Rational irrationality: Modeling climate change belief polarization using bayesian networks. *Topics in Cognitive Science*, 8(1):160–179.

Alice H. Eagly and William Revelle. 2022. Understanding the magnitude of psychological differences between women and men requires seeing the forest and the trees. *Perspectives on Psychological Science*, 17(5):1339–1358. PMID: 35532752.

Adam Enders, Casey Klofstad, and Joseph Uscinski. 2024. The relationship between conspiracy theory beliefs and political violence. *Harvard Kennedy School Misinformation Review*, 5(6).

Zach Horne, Derek Powell, John Hummel, and Keith Holyoak. 2015. Countering antivaccination attitudes. *Proceedings of the National Academy of Sciences of the United States of America*, 112.

Ana Ma and Derek Powell. 2025. Can large language models predict associations among human attitudes? In *Proceedings of the 47th Annual Meeting of the Cognitive Science Society*. Cognitive Science Society.

Keiichi Namikoshi, Alexandre L. S. Filipowicz, David A. Shamma, Rumen Iliev, Candice Hogan, and Nikos Aréchiga. 2024. Using llms to model the beliefs and preferences of targeted populations. *ArXiv*, abs/2403.20252.

Trisevgeni Papakonstantinou and Zachary Horne. 2023. Characteristics of persuasive deltaboard members on reddit'sr/changemyview.

Derek Powell, Kara Weisman, and Ellen Markman. 2022. Modeling and leveraging intuitive theories to improve vaccine attitudes.

Derek Powell, Kara Weisman, and Ellen M Markman. 2023. Modeling and leveraging intuitive theories to improve vaccine attitudes. *Journal of Experimental Psychology: General*.

Hunter Priniski and Keith Holyoak. 2020. Crowdsourcing to analyze belief systems underlying social issues.

Hunter Priniski and Zach Horne. 2019. Crowdsourcing effective educational interventions.

John Priniski and Zachary Horne. 2018. Attitude change on reddit's change my view. In *CogSci*.

Brittany Schotsch and Derek Powell. 2022. Understanding intuitive theories of climate change. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th international conference on world wide web*, pages 613–624.

RaiderBDev `stuck_in_the_matrix`, Watchful1. Reddit comments/submissions 2005-06 to 2024-06.

A Detailed overview of data pre-processing steps

A two-step filtering process was applied to derive the final dataset suitable for this study. Because posts on CMV span a wide variety of topics, we limited the dataset to a predefined set of topics in order to focus the analysis on clearly defined beliefs. Constraining the beliefs we focused on for this analysis, allowed us to define a ground truth dataset, which enabled a reliable validation of the analytical pipeline. To arrive at a reasonable selection of belief topics, we utilised the themes and questions from the Pew Research Centre surveys.

Pew Research Center, a nonpartisan American think tank that conducts public research on current issues, global attitudes and trends (see https://www.pewresearch.org), provides a reliable and widely recognised dataset on public opinion across various topics most relevant to the US population. We consulted the full surveys administered by Pew on various topics available at https://www.pewresearch.org/tools-and-resources/. To identify relevant topics and beliefs, two researchers (TP and AZ) independently coded all survey items in terms of (1) whether they represent a belief/attitude and (2) whether the belief is of a reasonable granularity level (i.e. overly specific items, such as "Do you think that all Americans should have the right to have data collected by law enforcement, such as criminal records or mugshots removed from public online search results?" were excluded or summarised within a broader item), resolving any discrepancies through discussion. The purpose of this filtering was to identify more general and common topics that would be prevalent in the Change My View dataset. Based on this coding, we arrived at a final set of 64 belief topics (e.g. "It is the government's responsibility to ensure its citizens have healthcare", "White people benefit a great deal from advantages in society that Black people do not have"). The full list of survey items and raw data are available at https://osf.io/smdt2/?view_only=e8aef33dcdad43f6a55cb29fec3a1745.

We followed a two-stage process to identify posts correlating with the belief topics identified from Pew Research. Initially, we filtered posts using a keyword set to capture posts relevant to each topic, optimised for relevance, but not exclusivity. In this step the aim was to capture all posts that might address the topic at hand without necessarily excluding noise, so we relied on a minimal set of search terms representing keywords that would almost certainly be used in a post referring to that topic. For example, we aimed to find posts relevant to the topic of healthcare so we started with a set of keywords "health", "insur", "healthcare", "cover", combining stemmed and unstemmed terms as appropriate and iteratively refined it until it returned posts obviously relevant to the topic. In the second stage of this retrieval process, we further refined the search by using GPT-4o to filter out noise captured by the hand search approach. Only posts by users with multiple contributions were retained. The resulting dataset consisted of 3,082 posts from 346 users after filtering.

The prompt used for the second stage of the filtering process read: 'Decide if this post is relevant to or partly or fully addresses the following belief topic: [topic]. If yes, the output of this request should be 'YES', if not it should be 'NO'. Please do not include anything else in the output. This is the post: [post title and text]'.

B Description of CMV Dataset

Table 2 is an example of a specific post, the extracted belief statements, property, and associated query and response options.

C Finetuning Procedure

We fine-tuned the GPT-4o model using OpenAI's fine-tuning API (https://platform.openai.com/docs/guides/fine-tuning). We used OpenAI's default fine-tuning hyperparameters, which include an Adam optimizer with an initial learning rate of $2e{-}5$, a batch size of 8, and up to 4 training epochs. No manual hyperparameter tuning was performed. All training and inference steps were conducted through OpenAI's managed infrastructure, and no additional modifications to the model architecture were made.

Phase 1: Belief Extraction For the first phase of analysis the finetuning system prompt reads: 'You are a cognitive scientist studying belief networks. You are trained in data annotation and can extract and list belief statements made from raw text from social media posts'. The user prompt read: 'You are given a post from Reddit where someone is expressing and justifying an attitude about a topic. Using the text of the post, you have to say what the main beliefs they have are, listing them as self-sufficient statements in bullet points without any

Original Post	Belief Statement(s)	Broader Belief (Theory)	Model	Human
I don't believe that there are any LOGICAL reasons why gay marriage shouldn't be allowed. CMV. All the people that are against gay marriage are against it because of moral or religious reasons, or because they feel that there isn't any point to gay marriage (e.g. they don't reproduce, similar to straight couples who decide to not have children). I still haven't seen a single logical reason why gay marriage shouldn't happen. I have no problem with gay marriage but when I try to understand the basis for the argument of the other side, I see people bashing it because of their personal beliefs and religious morals, not justified, fair thinking.	i don't believe that there are any logical reasons why gay marriage shouldn't be allowed; all the people that are against gay marriage are against it because of moral or religious reasons, or because they feel that there isn't any point to gay marriage (e.g., they don't reproduce, similar to straight couples who decide not to have children); there is not a single logical reason why gay marriage shouldn't happen.	Legalization of same-sex marriage has a positive impact on society.	4	5

Table 2: Example belief inference instance showing the original post, extracted belief statement, general theory, and ratings from the model and a human annotator.

other text, separated by commas. This is the post: [post title and text].'

Phase 2: Theory Inference For this phase, the system prompts remained consistent with Phase 1. User prompts were as follows: 'I am going to give you a set of specific statements that someone holds as beliefs. I will also give you a more general theory. I want you to calculate the likelihood on a scale of 1-7 that someone who holds the set of statements as beliefs, believes in the theory. Express that likelihood on a scale of 1-7, where 1 means that there is no evidence they hold the more general theory and 7 that they are extremely likely to hold the more general theory. This is the list of statements, between triple quotes: """[belief statements]""". This is the more general theory, between asterisks *[belief]*. The output should only contain the number of the likelihood and nothing else.'. Figure 1 presents a detailed overview of all data processing steps prior to LLM annotation at scale.

D Model Evaluation

The pipeline evaluation and comparison across models and against the human benchmark was conducted in R (version 4.4.1). The code used is available in the project repository https://osf.io/smdt2/?view_only=e8aef33dcdad43f6a55cb29fec3a1745.

Figure F presents a visual representation of the performance of the models against the human benchmark.

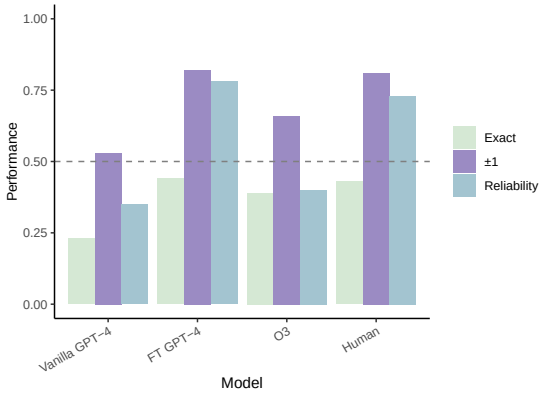

Figure 3: Evaluation of Baseline GPT-4o, FT GPT-4o, o3-mini, and human (inter-rater) performance across the metrics of exact agreement, ±1 agreement, and Cohen's κ

E Human Annotation Details

The annotators were trained using a detailed guideline to ensure consistency and accuracy in belief labelling. The guide followed by the annotators is available at https://osf.io/smdt2/?view_only=e8aef33dcdad43f6a55cb29fec3a1745.

To establish inter-annotator reliability for Phase 1, we evaluated the cosine similarity between the text extractions of the human annotators. To arrive at a similarity metric, the text entries were cleaned by removing punctuation, special characters, extra-

neous whitespace and turned all text to lowercase. Each cleaned text was then tokenised into individual words, and word counts per document were computed. Using these token counts, we generated term frequency-inverse document frequency vectors to represent each document's content in a weighted feature space. We then calculated pairwise cosine similarity scores between document vectors on matched pairs. Finally, we averaged them to obtain an overall similarity metric reflecting agreement between the two annotations.

This yielded a score of 0.63, indicating moderate agreement. Additionally, we conducted an 80/20 train-test split and evaluated cosine similarity between human annotations and GPT-4o outputs, resulting in a score of 0.57. While these scores indicate moderate reliability, it's important to note this level of agreement is quite notable given the inherent complexity of the task, and that the model's performance is similar to the discrepancy observed between human annotators before the triangulation discussion.

F Inferred Belief Correlations in CMV

Broader Belief (Theory)	N Posts	N Users
Corporate profits are too high and should be more reasonable	153	118
Demographics should not influence college admissions	215	193
It is the government's responsibility to ensure an adequate standard of living for its citizens	123	108
Marriage and children should be prioritized over other life choices	183	156
Ordinary people would do a better job solving the country's problems than elected officials	351	265
People convicted of crimes serve too little time in prison	172	136
The best way to ensure peace is through military strength	138	112
The economic system unfairly favors powerful interests	283	204
There are other countries that are better than the U.S.	185	159
White people benefit a great deal from advantages in society that Black people do not have	254	182

Table 3: Description of the CMV dataset used for comparison with the lab sample

The complete annotation and evaluation datasets are available in the project repository. Table 3 shows the number of associated posts and unique users posting under each pre-specified belief topic.

The correlation analysis revealed several notable relationships, with correlations of $\rho = .20 - .31$ being the most substantial. "Corporate profits too high" negatively correlates with "No demographics for college admissions" and positively correlates with both "Economic system unfairly favors the powerful" and "Other countries better than the U.S.". "No demographics for college admissions" shows a moderate positive correlation with "Marriage and children priority" and a moderate negative correlation with "Other countries better than the U.S.". "Standard of living government's responsibility" is moderately negatively correlated with "Convicts serve too little time", and "Marriage and children priority" also displays a moderate negative correlation with "Convicts serve too little time". These relationships represent the strongest correlations in the matrix, indicating that views on corporate profits, college admissions, government responsibility, and crime and punishment are more strongly related than other beliefs. Figure 2 present the matrix of correlations between each pair of beliefs.

Overall, these correlations suggest the presence of distinct clusters of beliefs in the dataset, that align with ideologically consistent patterns, such as those typically associated with liberal and conservative viewpoints. These findings support the validity of the method used, demonstrating that the relationships between beliefs are consistent with well-established patterns. The observed patterns provide further evidence that the underlying structure of the data reflects coherent ideological divisions.

G Human Data Sample

Survey data includes responses from 376 U.S. adults (223 male, 147 female, 4 non-binary, aged 18 to 73; Avg. age = 37.41, SD = 11.47) collected via Connect. Participants rated 64 beliefs on a 5-point Likert scale, consistent with Pew Research survey methodology (see ma.powell2025Can for more details).