Can Large Language Models Predict Associations Among Human Attitudes?

Ana Ma (yutongm1@asu.edu)
School of Social and Behavioral Sciences
Arizona State University

Derek Powell (dmpowell@asu.edu)
School of Social and Behavioral Sciences
Arizona State University

Abstract

Prior work has shown that large language models (LLMs) can predict human attitudes based on other attitudes, but this work has largely focused on predictions from highly similar and interrelated attitudes. In contrast, human attitudes are often strongly associated even across disparate and dissimilar topics. Using a novel dataset of human responses toward diverse attitude statements, we found that a frontier language model (GPT-4o) was able to recreate the pairwise correlations among individual attitudes and to predict individuals' attitudes from one another. Crucially, in an advance over prior work, we tested GPT-4o's ability to predict in the absence of surface-similarity between attitudes, finding that while surface similarity improves prediction accuracy, the model was still highly-capable of generating meaningful social inferences between dissimilar attitudes. Altogether, our findings indicate that LLMs capture crucial aspects of the deeper, latent structure of human belief systems.

Keywords: Artificial Intelligence; Psychology; Natural Language Processing; Social Cognitive

Whether about politics, ethics, science, sports, or the weather, people's beliefs and attitudes rarely stand alone. Instead, they are interrelated with one another, interwoven through cultural and social influences (Converse, 1964; Hofstede, 2001), and fit together as part of their intuitive theories for how the world works (Gerstenberg & Tenenbaum, 2017; Weisman & Markman, 2017; Powell, Weisman, & Markman, 2023). The ability to appreciate and anticipate these interrelations is a crucial target of human social reasoning and inference, as one generally only need to learn about a subset of a person's views to infer a great deal about the rest.

Researchers have now begun examining whether large language models (LLMs) might be able to predict or mimic the attitudes of different groups or individuals (Hwang, Majumder, & Tandon, 2023; Santurkar et al., 2023). Within the Artificial Intelligence and Natural Language Processing communities, this is sometimes likened to adopting the "persona" of group or individual users. This could be useful for aligning chat assistants to the cultural perspectives of their users (Choenni & Shutova, 2024; Pawar et al., 2024), for developing automated educational tools (Razafinirina, Dimbisoa, & Mahatody, 2024; Sonkar, Ni, Chaudhary, & Baraniuk, 2024; Tessler et al., 2024), for targeting recommendations or advertising (El-Sayed et al., 2024; Tang, Sun, Curran, Schaub, & Shin, 2024b) [1], or even for generating simulated human responses for social science research (Argyle et al., 2023).

A number of recent works have found that LLMs were able to predict human attitudes from demographic, ideological, and attitudinal measures (e.g. Argyle et al., 2023; Santurkar et al., 2023; Hwang et al., 2023; Long, Kawaguchi, Kan, & Chen, 2024). For instance, Hwang et al. (2023) found that prompting LLMs with a person's responses for a subset of related survey questions improved its ability to predict their other survey responses as compared to prompting with the respondents' demographics or ideology alone.

However, it remains unclear the extent to which these predictive capacities demonstrate LLMs' abilities to make meaningful inferences about relationships between human attitudes and opinions. Thus far, these models have only been tested in contexts where they have access to information about people's opinions for matters highly similar to the opinions-to-be-predicted. Therefore, it is possible that prior successes are driven largely by relatively surface-level semantic similarities between the prompted and predicted statements, rather than by knowledge of deeper latent connections between attitudes.

Conversely, a striking feature of human attitudes are the often strong interrelations among ostensibly distant attitudes or beliefs. Perhaps most obviously, people's views about political, religious, or moral concerns tend to be organized by broader ideologies. For instance, in the U.S., conservatives tend to favor strong restrictions on abortion while favoring little to no restriction on gun sales. Meanwhile, many other beliefs are related due to their intrinsic interconnections. For instance, Powell et al. (2023) examined intuitive theories surrounding vaccination decisions. They found that people's "vaccination intentions" were most strongly related to highly proximal beliefs, such as about "vaccine danger" and "vaccine effectiveness". However, they also found quite strong relationships between vaccination intentions and more distant beliefs, such as about the merits of natural versus artificial things, and about the role of "balance" in determining health.

As these examples illustrate, relations among people's beliefs and attitudes can be non-obvious—e.g., gun control and reproductive rights have nothing in common on their face—and therefore unlikely to be predictable from the semantic similarity of statements of the beliefs themselves. Rather, appreciating these interrelations requires a deeper understanding of both social and epistemic concerns.

[1] These uses present substantial risks and potential ethical concerns, as discussed in the works we cite here.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

Interrelations among human attitudes and beliefs

A *belief system* consists of a grouping of human attitudes that can form via a range of constraints (Converse, 1964). One constraint is ideology: for example, conservatism is characterized by resistance to change and a preference for tradition (Adorno, Frenkel-Brunswik, Levinson, & Sanford, 1950), and is associated with both political attitudes (e.g., against legalized abortion) as well as non-political views (e.g., skepticism of modern art; Wilson & Patterson, 1968). Moreover, ideologies themselves are frequently correlated with one another (Jost, Glaser, Kruglanski, & Sulloway, 2003). For example, correlations have been found among political conservatism, Social Dominance Orientation (i.e., support for hierarchical social structures; Pratto, 1999; Altemeyer, 1998), and System Justification Theory (i.e., legitimizing existing systems; Jost, Pelham, Sheldon, & Sullivan, 2003). Through the linkages among ideologies, an even broader web of attitudes can be drawn together into a system of beliefs.

Social influences are another source of constraint on cohesive belief patterns. Social ties promote interactions that reward similarity and alignment in behaviors (Schachter, 1959) and attitudes (Byrne, 1961; Moussaïd, Kämmer, Analytis, & Neth, 2013). Individuals' come to internalize the social norms they experience across familial, educational, religious, and cultural environments, producing bundles of related beliefs (Converse, 1964). For example, a community with individualistic norms can establish attitudes about work (e.g., valuing personal achievement over collective gain), romantic relationships (e.g., emphasizing personal choice over arranged relationships), philosophy of life (e.g., focusing on finding one's own path over fulfilling a societal role), and other concerns (Hofstede, 2001). Not only are individuals passively influenced by these social norms, but they also further actively self-select into groups of shared views. This tendency toward "homophily" (McPherson, Smith-Lovin, & Cook, 2001), creates feedback-loops that further reinforce associations among attitudes (Axelrod, Daymude, & Forrest, 2021).

Of course, perhaps the most fundamental constraint on beliefs and attitudes is *coherence* (Thagard, 1989): attitudes and beliefs are connected with one another through their meanings, through logical implication, and through people's understanding of causal relationships in the world (Gerstenberg & Tenenbaum, 2017; Powell et al., 2023). In the face of new evidence, people attempt to update their beliefs to maintain coherence (Holyoak & Simon, 1999; Spellman, Ullman, & Holyoak, 1993), and a clash between beliefs can lead to the discomfort of "cognitive dissonance" (Festinger, 1957).

Social inference and LLM personas

OpinionQA (Santurkar et al., 2023) has emerged as a leading benchmark for researchers and developers aiming to align LLMs to users along demographic, ideological, and attitudinal dimensions. This benchmark leverages high-quality survey data collected by Pew from several representative samples of U.S. respondents to provide both categorical and individual-level persona information. The benchmark is composed of 1506 survey questions and answers from 80,098 respondents measured across 15 American Trends panel surveys conducted by Pew. Each American Trends panel survey focuses on a set of topics of concern for American civil and political life.

A number of prior works have focused on aligning LLM responses with specific groups of users (Santurkar et al., 2023). Going further, Hwang et al. (2023) examined more fine-grained relationships between specific attitudes. They found that prompting an LLM with a person's responses to other specific attitude questions substantially improved the model's prediction of the target attitude. Moreover, they found that selecting a subset of attitudes for prompting improved predictions: filtering to choose the top-k (3 or 8) attitude questions most similar to the target statement (in terms of cosine similarity) improved predictions as compared with prompting with all responses.

Due to the structure of the Pew panel data on which the OpinionQA benchmark is based, survey respondents tended to be asked about a number of similar questions. As such, Hwang et al. (2023)'s methodology for selecting the top-k most similar items and responses as the language model prompt will tend to produce highly-similar prompting sets, which might allow for successful prediction from a relatively simple inference processes. Roughly, good predictions might be achieved through a heuristic process of similarity-matching, something like: predict the response given to the most similar item in the prompt. Subsequent works have found that using a more sophisticated process to select prompt information can produce more accurate predictions (Long et al., 2024), suggesting that modern LLMs engage in inferential processing beyond simple similarity-matching heuristics. However, this work did not examined this question directly.

Another line of work has examined the potential application of LLMs in simulating human responses for social science research (Argyle et al., 2023; Sun et al., 2024). To validate the viability of this application, Argyle et al. (2023) define and evaluate several criteria for establishing algorithmic fidelity to human responses. Most important for our question is what they call "Pattern Correspondence": whether correlations among attitudes from silica samples reflect those in human samples. As in other work using OpinionQA (Hwang et al., 2023; Long et al., 2024), their findings may owe to semantic similarity of the items measured.

The present studies

We sought to examine the the extent to which frontier language models can predict peoples' attitudes from one another in the absence of direct semantic similarity between the prompted and target attitudes. To support this novel examination across more distant connections, we surveyed a sample of U.S. respondents on a wide-ranging subset opinion questions from the Pew surveys underlying OpinionQA (Santurkar et al., 2023). Then, we tested the ability of OpenAI's GPT-4o

(OpenAI et al., 2024) to predict individual attitudes on the basis of their other survey responses, for both semantically-similar and semantically-dissimilar items.

Across our analyses, we find that GPT-4o imperfectly but meaningfully recreates the observed pairwise correlations among attitudes and predicted individual attitudes. Semantic similarity between attitude statements played a limited role in this capability: GPT-4o's predictions were biased by similarity and were more accurate when made based on semantically-similar items, but this influence was modest. Our results clearly demonstrate that GPT-4o is able to predict individuals' attitudes in the absence of surface-level semantic similarity between prompted and targeted attitude statements. Altogether, our findings indicate that frontier language models like GPT-4o encode some deeper knowledge of latent associations among human attitudes.

Study 1: Human Data Collection

Participants

A sample of U.S. adults were invited to participate in the survey via Connect with a compensation of $2.25 for their time.[2] After excluding 10 participants who failed a simple attention-check, 376 participants' (223 male, 147 female, 4 non-binary; 18 to 73 years old: Avg. age = 37.41, SD = 11.47) responses were included in our study. Participants reported a variety of races (237 White or Caucasian, 72 Black or African American, 29 Asian, 17 Hispanic or Latinx, 14 Multiracial/Biracial, 1 Native American or Alaskan Native, 1 none of the listed above, and 5 preferred not to say) and political backgrounds (198 in the Democratic Party, 85 Republican, 80 Independent, 6 Other, and 7 preferred not to say).

Materials and Procedure

Drawing on the opinion questions in OPINIONQA, we selected a subset of 64 diverse items assessing views on a wide range of topics relevant to the U.S. society. We transformed these questions from Pew into declarative statements to allow responses on a common agree-disagree scale for all items.

Participants were surveyed in an online Qualtrics survey. After consenting to participate, participants rated their agreement with each statement on a five-point scale, ranging from *Strongly agree* to *Strongly disagree*. All participants were asked to respond to all 64 statements in a random order. Two attention checks were evenly-spaced in the study.

Results

Pearson correlation coefficients were calculated between all statement pairs (n = 2,016). As shown in Figure 1, substantial associations were found among attitudes both within and across topic areas. As an illustrative example, participants' agreement level between statements "The government should prioritize addressing climate change." (*Environment and Climate*) and "Increasing the number of guns is bad for society." (*Public Safety and Security*) were positively correlated $r = 0.58$. Many other correlated attitudes were found between distinct social topics, as represented by non-diagonal colored tiles in Figure 1.

Study 2: Estimating correlations with GPT-4o

Next, we tested whether predictions from GPT-4o reflect the correlational patterns observed between human responses to these 64 atttiude questions.

Model prompting

To prompt GPT-4o to make its predictions, we created user prompts following the general structure used by Hwang et al. (2023). As shown in Figure 2, each prompt contained 1) instructions for the model to "Help predict a person's answers on a social attitudes survey"; 2) an example statement and answer; 3) instructions to generate a prediction based on the example statement-answer pair; and 4) a target statement and answer choices for the model to predict. For every possible unique pairing of the 64 attitude statements we created five prompts, with one statement as the target and the other serving as the "example" with each of the 5 possible answer choices. This produced 64 target statements $\times$ 63 example statements $\times$ 5 example answers $= 20,160$ total prompts. We then passed these prompts to GPT-4o through the OpenAI API and recorded its responses.[3]

Metrics

GPT-4o-estimated correlations The prompted agreement level and the corresponding GPT-4o-predict agreement levels were converted to a numeric score from 1 (*Strongly disagree*) to 5 (*Strongly agree*). We then calculated the Pearson correlations between GPT-4o-predicted and prompted agreement scores for each pair of example and estimated statements (n = 2,016).

Similarity calculation To examine the relationship between model response correlations and statement semantic similarities, we first collected the vector embeddings for all 64 statements using the OpenAI API, `text-embedding-3-large` embedding model. We then calculated the cosine similarities between all statement pairs' (n = 2,016) vector embeddings.

$$S_C(\mathbf{A}, \mathbf{B}) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \cdot \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Results

Human and GPT-4o-estimated correlations As shown in Figure 3, GPT-4o was able to estimate human attitudes by capturing the inter-correlations among those attitudes. Comparing the observed correlations among human responses and the correlations estimated from GPT-4o's predictions, we found these coefficients to be themselves strongly and positively correlated with one another (r = .80, 95% CI

[2] All data and analyses can be found at https://github.com/derekpowell/social-llm

[3] In all cases we use version `gpt-4o-2024-08-06`, sampling up to 10 tokens with temperature 0.01 for consistency.

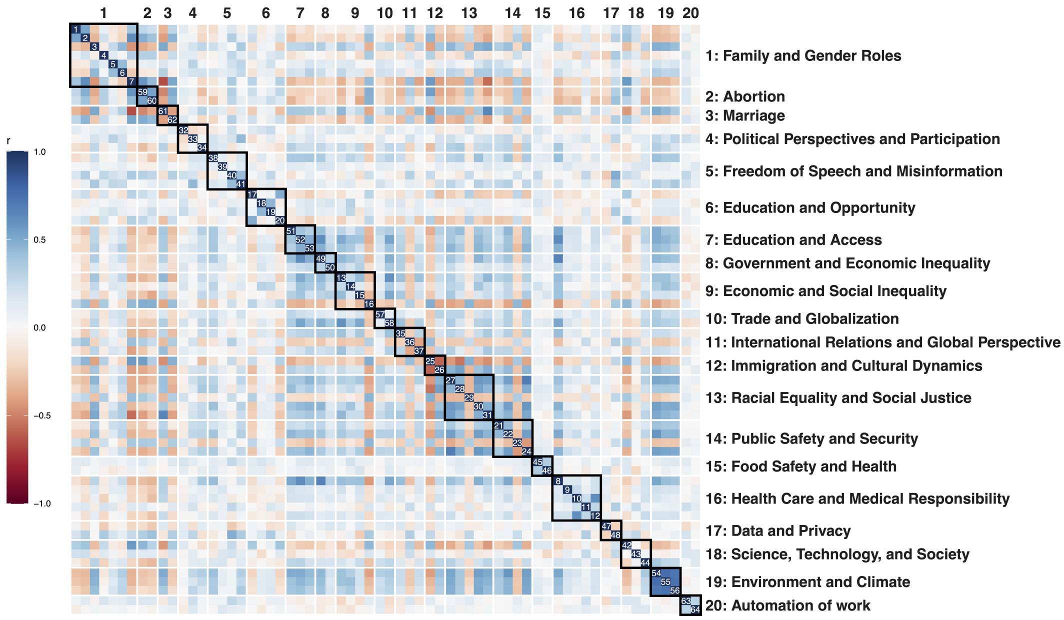

Figure 1: Correlations among self-reported human attitudes. Attitudes (labeled along the diagonal) are ordered by topic area (labeled on axes), and topic areas are separated by grid lines. Black bounding boxes highlight topic groupings along the diagonal indicating within-topic associations.

$[.78, .82]$, $p < .001$). However, whereas human attitude correlations were approximately normally distributed from -0.65 to 0.76, correlations estimated from GPT-4o tended to be more extreme, following a bimodal distribution with peaks near -1 and 1. Figure 3 illustrates this pattern in the clustering of points at the extremes along the x-axis.

Semantic Similarity and Estimated Correlations Next, we examined the extent to which GPT-4o's predicted associations between attitudes rely on the semantic similarity between those attitudes. Cosine similarity of the statements was very weakly correlated with the absolute GPT-estimated strength of association among statements ($r = .12$, 95% CI $[.08, .16]$, $p < .001$). This relationship is quite imperfect, so that dissimilar statements sometimes have strong associations, and similar statements sometimes have only weak associations.

As GPT-4o reliably overestimates the strength of associations, in the following analyses we rank-transform both human and LLM-predicted correlations for comparison purposes. We do find some evidence that similarity biases GPT-4o's correlation estimates: Regressing the absolute human correlation values on cosine similarity for each pair, we divide attitude pairs into those more and less-strongly associated than would be predicted by cosine similarity. We find that GPT-4o more-commonly overestimates the association for pairs that are less-strongly associated than predicted from cosine similarity (64.0% llm-based-rank > human-rank) and underestimates associations for pairs that are more-strongly associated (72.1% llm-based-rank < human-rank).

Nevertheless, the biases induced by semantic similarity appear to have only minor impacts on GPT-4o's ability to recapitulate correlations among attitudes. Filtering for only dissimilar (i.e. $S_c < .20$) attitude pairs significantly reduces the correlation between GPT-4o-estimated correlations and human correlations ($p = .001$), yet the relationship between estimated and observed correlations remained strong ($r = .77$).

Study 3: Predicting human responses using GPT-4o

Finally, we examined GPT-4o's ability to predict individual respondents' answers to each target attitude from their responses to other attitude questions. Crucially, we are interested in examining the degree to which GPT-4o might rely upon the semantic similarity of the target attitude and the attitudes in its prompt.

Selecting and evaluating top-k predictors

To test this, we examined GPT-4o's ability to predict target attitudes using two feature selection methods. First, following Hwang et al. (2023), we identified a set of *semantically-similar* predictors for each target attitude by selecting the top-k most similar attitudes based on cosine similarity (with $k \in \{3, 8\}$). For each target, we also identified a set of *semantically-dissimilar* predictors. Here, we exclude any predictors with $S_c > .20$ and from the remainder choose the top-k most correlated attitudes based on human data.

Help predict a person's answers on a social attitudes survey. The person was asked the following question and gave the following answer:

Question: Please rate the extent of your agreement or disagreement with the following statement:
Statement: The government should prioritize addressing climate change.
Answer: Strongly agree

Based on the above information, which answer choice will this person select for this question:

Question: Please rate the extent of your agreement or disagreement with the following statement:
Statement: Increasing the number of guns is bad for society.
Answer choices: Strongly agree; Somewhat agree; Neither agree nor disagree; Somewhat disagree; Strongly disagree

Respond only with the answer choice you predict the person would give.

Figure 2: Example prompt used to predict human response to a target statement based on an example statement-answer pair.

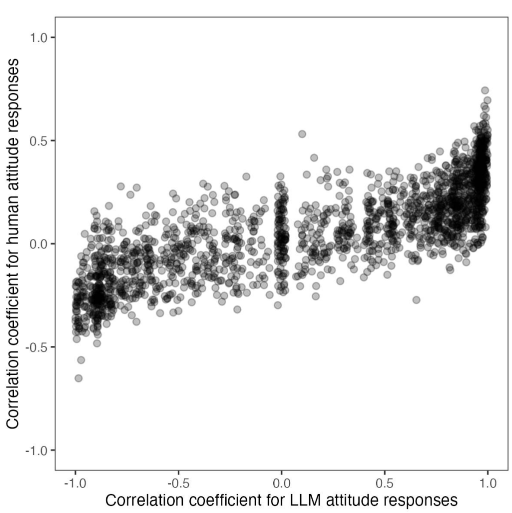

Figure 3: Scatterplot of correlation values observed in human data against correlation values among GPT-4o estimations. Points are jittered on the x-axis for better visibility.

"Oracle" model training

For each target attitude and set of semantically-similar and dissimilar predictors, we trained a random forest model to examine how well those predictors could in-fact predict the target item. We evaluate the model fit using a 10-fold cross validation procedure, evaluating predictive accuracy using the model predictions on the held-out validation splits. We take these models as "oracle" models and we take their validation performance to represent the upper-bound of predictability for the target attitudes from the selected predictors.[4]

Generating GPT-4o predictions

To predict each participant's responses, we constructed prompts containing 1) instructions for the model to "Help predict a person's answers on a social attitudes survey"; 2) a set of k (i.e., 3 or 8) example statements and a participant's actual responses observed in Study 1; 3) instructions to generate a prediction based on those statement-answer pairs; and 4) the target statement along with the list of answer choices (i.e., the five agreement levels). For each item, we composed predictor sets for the top-3 and top-8 items chosen using our criteria for selecting semantically-similar and semantically-dissimilar items. We then passed the prompts to GPT-4o through the OpenAI API.

Metrics

Accuracy For GPT-4o, we define accuracy as the exact match between the model's output and the choice text.

Results

Table 1 shows the predictive accuracy of "oracle" models and GPT-4o based on the top-k semantically-similar and semantically-dissimilar item selection strategies. As shown by the oracle model results, the semantically-similar items were in-fact more predictive of the target attitudes. This is to be expected: semantically-similar items often have relatively strong correlations with the target, so that excluding these items puts the dissimilar models at a substantial disadvantage, despite choosing the dissimilar items based on observed correlations. When including all items, GPT-4o's predictions were more accurate when prompted with similar items.

Although the semantically-similar items were more predictive overall, for a substantial portion of target attitudes ($24/64$ for $k = 3$ and $14/64$ for $k = 8$), oracle models using semantically-dissimilar items were superior to those using similar items. Again this is generally to be expected: the most semantically-similar items are not *always* those that are most predictive. Importantly, these cases offer strong a test of the role of semantic similarity in informing language models' attitude predictions: If the language model predicts as well or better with low-similarity items, this would then demonstrate that it is not relying on simple semantic similarity. However, if its predictions are worse in these cases where they could

[4]We qualify this label with a potential limitation: given that these models are estimated from a limited sample of human data, superior predictive accuracy may technically be possible.

Table 1: Predictive accuracies of GPT-4o and "Oracle" Random Forest models.

Model	Target subset	Selection method	Top-k	Accuracy
GPT4o	all	similar (S_c)	3	43.5%
			8	45.2%
		dissimilar (r)	3	40.4%
			8	41.9%
GPT4o	dissimilar > similar	similar (S_c)	3	41.2%
			8	40.8%
		dissimilar (r)	3	42.0%
			8	42.2%
Oracle	all	similar (S_c)	3	47.7%
			8	49.6%
		dissimilar (r)	3	44.4%
			8	45.5%
Oracle	dissimilar > similar	similar (S_c)	3	43.9%
			8	45.1%
		dissimilar (r)	3	47.3%
			8	47.6%
Chance	all	–	–	29.4%

Note: Models are evaluated against all observed responses from all participants (64 items $\times$ 376 participants $= 24,064$).

or should be better, then we can conclude it is in some way relying on semantic similarity.

Focusing on just these items, the results in Table 1 demonstrate that GPT-4o is also more accurate when predicting from the dissimilar items than similar items. We take this as a clear demonstration that GPT-4o is capable of making meaningful social inferences in the absence of semantic similarity between statements. At the same time, comparing the accuracy of these two predictor sets between GPT-4o and the oracle models fit to human data reveals that GPT-4o's predictions are closer to the upper-bound of predictability when it is prompted with semantically-similar items (by 2.6% for $k = 3$ and 1.1% for $k = 8$). This suggests that GPT-4o does, at the same time, leverage the more surface-level similarity between attitudes in generating predictions.

Discussion

Using a novel dataset of human attitude across a range of social topics, we tested the ability of a frontier language model (i.e. GPT-4o) to estimate the correlations among human attitudes and to predict individuals' attitudes from one another. Our findings demonstrate that GPT-4o can largely replicate the inter-correlations observed in human data, and can meaningfully predict individuals' attitudes. Going beyond prior findings (Hwang et al., 2023), we show that LLMs attitude predictions are not merely driven by surface-level semantic similarity, but instead tap in to latent associations among human attitudes.

We must note several limitations of our studies which might point toward directions for future work. First, we only calculated model estimation alignment and prediction accuracy based on five levels of outputs for each target-example statement pair. Although we experimented with repeated sampling and prompting for probability distribution outputs, these approaches did not result in much variability or improvement in model estimations.

Second, we only prompted GPT-4o to respond immediately with a predicted response option. This may underestimate the potential performance of the model, and future research could explore other prompting strategies. For instance, model prompts could utilize the chain-of-thought methods (CoT) (Wei et al., 2023) to encourage step-by-step "thinking" from the model, which might improve performance (Long et al., 2024). Alternately, future work might examine the capabilities of new reasoning models such as O1 (OpenAI, 2024) and DeepSeek R1 (DeepSeek-AI et al., 2025). These approaches might improve performance or decrease the biases we observed with respect to semantic similarity. Further, examination of the intermediate CoT or reasoning steps might support deeper insights into the information leveraged by the model. For instance, we might examine whether models perform better when they explicitly infer (and tokenize in their reasoning) the ideology of a target person, or when they take other latent dimensions of attitudes into account.

Our findings underscore a number of existing concerns about the safe deployment of LLMs and other AI systems. First, LLM alignment efforts seeking to personalize models to users' existing attitudes risk creating echo chambers in human-AI interactions, with the potential to increase polarization (Axelrod et al., 2021). In addition, the major objective of monetizing LLM chatbots has driven large tech companies to adopt models for advertisement creation and promotion, e.g. Microsoft's Copilot (Microsoft, 2024). When AI ads are difficult to notice even with ad disclosures, users reported feeling manipulated, intruded, and less trusting towards LLM chatbots (Tang, Sun, Curran, Schaub, & Shin, 2024a).

Moreover, AI safety advocates and researchers have identified a number of capabilities that, if achieved by AI systems, could present substantial risk to humans. Among these are the ability to persuade or manipulate humans (Ji et al., 2024; Burtell & Woodside, 2023). Persuasion often benefits from understanding of others' existing viewpoints (Lewandowsky, Ecker, Seifert, Schwarz, & Cook, 2012; Powell et al., 2023), suggesting that the ability to predict attitudes may be a key aspect of this larger capacity. Our findings therefore underscore existing concerns about the potential risks of AI systems for persuasion and manipulation. Although there may be positive uses of such capabilities (Karinshak, Liu, Park, & Hancock, 2023), these capabilities nevertheless present alarming risks for automated propaganda and targeted misinformation.

Conclusion

Using a novel dataset of human responses to diverse attitude measures, we demonstrated the ability of frontier language models to replicate the interrelations among human beliefs without relying on surface-level semantic similarities. Our results suggest these models possess capabilities for higher-level social inference beyond mere pattern matching.

References

Adorno, T. W., Frenkel-Brunswik, E., Levinson, D. J., & Sanford, N. R. (1950). *The authoritarian personality*. New York, NY: Harpers.

Altemeyer, B. (1998). The other “authoritarian personality”. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 30, pp. 47–92). Academic Press. doi: 10.1016/S0065-2601(08)60382-2

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023, February). Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, 1–15. doi: 10.1017/pan.2023.2

Axelrod, R., Daymude, J. J., & Forrest, S. (2021). Preventing extreme polarization of political attitudes. *Proceedings of the National Academy of Sciences*, *118*(50), e2102139118. Retrieved from `https://doi.org/10.1073/pnas.2102139118` doi: 10.1073/pnas.2102139118

Burtell, M., & Woodside, T. (2023, March). *Artificial Influence: An Analysis Of AI-Driven Persuasion* (No. arXiv:2303.08721). arXiv. doi: 10.48550/arXiv.2303.08721

Byrne, D. (1961). Interpersonal attraction and attitude similarity. *The Journal of Abnormal and Social Psychology*, *62*(3), 713–715. doi: 10.1037/h0044721

Choenni, R., & Shutova, E. (2024, August). *Self-Alignment: Improving Alignment of Cultural Values in LLMs via In-Context Learning* (No. arXiv:2408.16482). arXiv. doi: 10.48550/arXiv.2408.16482

Converse, P. E. (1964). The nature of belief systems in mass publics (1964). *Critical Review*, *18*(1–3), 1–74. doi: 10.1080/08913810608443650

DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., … Zhang, Z. (2025). *Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning* (No. arXiv:2501.12948). Retrieved from `https://doi.org/10.48550/arXiv.2501.12948` (arXiv preprint) doi: 10.48550/arXiv.2501.12948

El-Sayed, S., Akbulut, C., McCroskery, A., Keeling, G., Kenton, Z., Jalan, Z., … Brown, S. (2024, April). *A Mechanism-Based Approach to Mitigating Harms from Persuasive Generative AI* (No. arXiv:2404.15058). arXiv. doi: 10.48550/arXiv.2404.15058

Festinger, L. (1957). *A Theory of Cognitive Dissonance*. Stanford University Press. Paperback.

Gerstenberg, T., & Tenenbaum, J. B. (2017). Intuitive theories. In *The Oxford handbook of causal reasoning* (pp. 515–547). New York, NY, US: Oxford University Press. doi: 10.1093/oxfordhb/9780199399550.001.0001

Hofstede, G. (2001). *Culture’s consequences: Comparing values, behaviors, institutions, and organizations across nations*. Thousand Oaks, CA: Sage Publications. doi: 10.1016/S0005-7967(02)00184-5

Holyoak, K. J., & Simon, D. (1999). Bidirectional Reasoning in Decision Making by Constraint Satisfaction. *Journal of Experimental Psychology: General*, 29.

Hwang, E., Majumder, B. P., & Tandon, N. (2023, May). *Aligning Language Models to User Opinions* (No. arXiv:2305.14929). arXiv.

Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., Wang, K., … Gao, W. (2024, May). *AI Alignment: A Comprehensive Survey* (No. arXiv:2310.19852). arXiv. doi: 10.48550/arXiv.2310.19852

Jost, J. T., Glaser, J., Kruglanski, A. W., & Sulloway, F. J. (2003). Political conservatism as motivated social cognition. *Psychological Bulletin*, *129*(3), 339–375. doi: 10.1037/0033-2909.129.3.339

Jost, J. T., Pelham, B. W., Sheldon, O., & Sullivan, B. (2003). Social inequality and the reduction of ideological dissonance on behalf of the system: Evidence of enhanced system justification among the disadvantaged. *European Journal of Social Psychology*, *33*, 13–36. doi: 10.1002/ejsp.127

Karinshak, E., Liu, S. X., Park, J. S., & Hancock, J. T. (2023). Working with ai to persuade: Examining a large language model’s ability to generate pro-vaccination messages. *Proceedings of the ACM on Human-Computer Interaction*, *7*(CSCW1), 1–29. doi: 10.1145/3579592

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012, December). Misinformation and Its Correction: Continued Influence and Successful Debiasing. *Psychological Science in the Public Interest*, *13*(3), 106–131. doi: 10.1177/1529100612451018

Long, D. X., Kawaguchi, K., Kan, M.-Y., & Chen, N. F. (2024, December). *Aligning Large Language Models with Human Opinions through Persona Selection and Value–Belief–Norm Reasoning* (No. arXiv:2311.08385). arXiv. doi: 10.48550/arXiv.2311.08385

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, *27*, 415–444. Retrieved from `https://www.jstor.org/stable/2678628` doi: 10.1146/annurev.soc.27.1.415

Microsoft. (2024). *Microsoft privacy statement – microsoft privacy.*

Moussaïd, M., Kämmer, J. E., Analytis, P. P., & Neth, H. (2013). Social influence and the collective dynamics of opinion formation. *PLOS ONE*, *8*(11), e78433. doi: 10.1371/journal.pone.0078433

OpenAI. (2024). *Introducing openai o1*. Retrieved from `https://openai.com/o1`

OpenAI, Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., … Malkov, Y. (2024, October). *GPT-4o System Card* (No. arXiv:2410.21276). arXiv. doi: 10.48550/arXiv.2410.21276

Pawar, S., Park, J., Jin, J., Arora, A., Myung, J., Yadav, S., … Augenstein, I. (2024, October). *Survey of Cultural Awareness in Language Models: Text and Beyond* (No. arXiv:2411.00860). arXiv. doi: 10.48550/arXiv.2411.00860

Powell, D., Weisman, K., & Markman, E. M. (2023, May). Modeling and leveraging intuitive theories to improve vaccine attitudes. *Journal of Experimental Psychology: General*, *152*(5), 1379–1395. doi: 10.1037/xge0001324

Pratto, F. (1999). The puzzle of continuing group inequality: Piecing together psychological, social, and cultural forces in social dominance theory. *Advances in Experimental Social Psychology*, *31*, 191–263.

Razafinirina, M. A., Dimbisoa, W. G., & Mahatody, T. (2024). Pedagogical Alignment of Large Language Models (LLM) for Personalized Learning: A Survey, Trends and Challenges. *Journal of Intelligent Learning Systems and Applications*, *16*(04), 448–480. doi: 10.4236/jilsa.2024.164023

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023, March). *Whose Opinions Do Language Models Reflect?* (No. arXiv:2303.17548). arXiv.

Schachter, S. (1959). *The psychology of affiliation*. Stanford, CA: Stanford University Press.

Sonkar, S., Ni, K., Chaudhary, S., & Baraniuk, R. G. (2024, October). *Pedagogical Alignment of Large Language Models* (No. arXiv:2402.05000). arXiv. doi: 10.48550/arXiv.2402.05000

Spellman, B. A., Ullman, J. B., & Holyoak, K. J. (1993, January). A Coherence Model of Cognitive Consistency: Dynamics of Attitude Change During the Persian Gulf War. *Journal of Social Issues*, *49*(4), 147–165. doi: 10.1111/j.1540-4560.1993.tb01185.x

Sun, S., Lee, E., Nan, D., Zhao, X., Lee, W., Jansen, B. J., & Kim, J. H. (2024, February). *Random Silicon Sampling: Simulating Human Sub-Population Opinion Using a Large Language Model Based on Group-Level Demographic Information* (No. arXiv:2402.18144). arXiv. doi: 10.48550/arXiv.2402.18144

Tang, B. J., Sun, K., Curran, N. T., Schaub, F., & Shin, K. G. (2024a). Genai advertising: Risks of personalizing ads with llms. *arXiv*, *2409.15436*. doi: 10.48550/arXiv.2409.15436

Tang, B. J., Sun, K., Curran, N. T., Schaub, F., & Shin, K. G. (2024b, September). *GenAI Advertising: Risks of Personalizing Ads with LLMs* (No. arXiv:2409.15436). arXiv. doi: 10.48550/arXiv.2409.15436

Tessler, M. H., Bakker, M. A., Jarrett, D., Sheahan, H., Chadwick, M. J., Koster, R., … Summerfield, C. (2024). Ai can help humans find common ground in democratic deliberation. *Science*, *386*(6719), eadq2852. doi: 10.1126/science.adq2852

Thagard, P. (1989, September). Explanatory coherence. *Behavioral and Brain Sciences*, *12*(3), 435–467. doi: 10.1017/S0140525X00057046

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., … Zhou, D. (2023). Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*. Retrieved from `https://doi.org/10.48550/arXiv.2201.11903` doi: 10.48550/arXiv.2201.11903

Weisman, K., & Markman, E. M. (2017, October). Theory-based explanation as intervention. *Psychonomic Bulletin & Review*, *24*(5), 1555–1562. doi: 10.3758/s13423-016-1207-2

Wilson, G. D., & Patterson, J. R. (1968). A new measure of conservatism. *British Journal of Social and Clinical Psychology*, *8*, 264–269.