Beyond Demographics: Aligning Role-playing LLM-based Agents Using Human Belief Networks

Yun-Shiuan Chuang **Krirk Nirunwiroj**[†] **Zach Studdiford**[†] **Agam Goyal**
Vincent V. Frigo **Sijia Yang** **Dhavan Shah** **Junjie Hu** **Timothy T. Rogers**
University of Wisconsin-Madison
{yunshiuan.chuang, nirunwiroj, studdiford, agoyal25}@wisc.edu
{vfrigo, syang84, dshah, junjie.hu, ttrogers}@wisc.edu

Abstract

Creating human-like large language model (LLM) agents is crucial for faithful social simulation. Having LLMs role-play based on demographic information sometimes improves human likeness but often does not. This study assessed whether LLM alignment with human behavior can be improved by integrating information from empirically-derived human belief networks. Using data from a human survey, we estimated a belief network encompassing 64 topics loading on nine non-overlapping latent factors. We then seeded LLM-based agents with an opinion on one topic, and assessed the alignment of its expressed opinions on remaining test topics with corresponding human data. Role-playing based on demographic information alone did not align LLM and human opinions, but seeding the agent with a single belief greatly improved alignment for topics related in the belief network, and not for topics outside the network. These results suggest a novel path for human-LLM belief alignment in work seeking to simulate and understand patterns of belief distributions in society.

1 Introduction

With rapid advances in large language models (LLMs), there has grown increasing interest in using these technologies to simulate human opinions [18, 17, 7, 23]. Contemporary LLMs can be prompted to role-play as individuals with particular demographic traits, sometimes then producing patterns of behavior that seem human-like. For instance, when asked to report the US unemployment rate when President Obama left office, ChatGPT will provide the exact answer; but if first instructed to role-play as a typical Democrat or Republican and asked the same question, the model produces incorrect, inflated estimates that mirror patterns of partisan bias in analogous human studies [8]. Such results raise the possibility that, with strategic prompting, LLMs may serve as useful proxies for capturing beliefs and attitudes of various socio-demographic groups.

Other recent work suggests, however, that the alignment between beliefs expressed by role-playing LLMs and matched human participants is unreliable at best. For instance, LLMs tuned via human feedback generally reflect opinions from liberal and well-educated demographics and that having LLMs role-play as humans with different socio-demographic traits or initial belief does not remediate this tendency [20, 22, 7]. Overall, this work suggests that LLM fine-tuned with human feedback tend to adopt progressive stances regardless of the demographic background they role-play–a behavior that may aid LLM safety, but limits their utility as models of human communicative dynamics.

The current paper considers an alternative approach to aligning the attitudes expressed by role-playing LLMs and the human groups they are intended to emulate. The central idea relies on behavioral

[†]Joint second authors.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

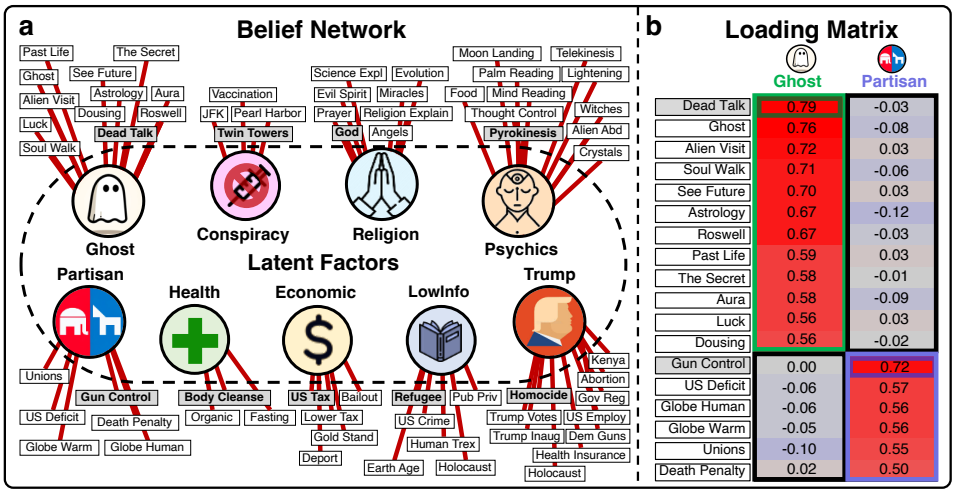

Figure 1: (a) The belief networks estimated by factor analysis from human respondents' responses on the Controversial Beliefs Survey. The nine central nodes are the orthogonal latent factors, and the leaves (rectangles) are the 64 individual topics x. The training topics x_{train} are highlighted with grey backgrounds. (b) Factor loading matrix between two latent factors and their topics. Figure 4 shows the full factor loading matrix.

studies of human *belief networks*: the empirical observation that beliefs on different topics tend to cohere together in patterns of high-order covariation [3, 26, 15, 25]. For instance, people who believe that government should support social welfare programs are also more likely to believe in higher taxes on the wealthy, strong union protections, and universal health care. Thus, knowing a person's opinion on one topic can carry rich information about their likely views on many others. Because LLMs learn from vast amounts of human-generated language, the weights they acquire and hence patterns of behaviors they exhibit may implicitly capture the tendency for various beliefs to co-occur in human populations, providing novel leverage for alignment. Specifically, human-LLM alignment may be guided, not just by socio-demographic role-playing, but also by instructing the LLM to hold a specific opinion on a representative topic. See Appendix A for more related work on belief networks and human-LLM opinion alignment.

To test this idea, we considered a belief network constructed in prior work which applied factor analysis to a dataset measuring human beliefs across a diverse array of topics [11]. Factor analysis decomposes patterns of covariation among expressed beliefs, identifying relationships between beliefs and a set of underlying latent factors. This factor analysis identified nine latent factors. For example, the *ghost factor* groups beliefs in various supernatural phenomena (e.g., talking to the dead). We then considered how well the opinions of LLMs align with human participants when prompted with and without information extracted from the estimated belief network. The results suggest that attention to empirically-derived human belief networks provides a useful strategy for human-LLM alignment.

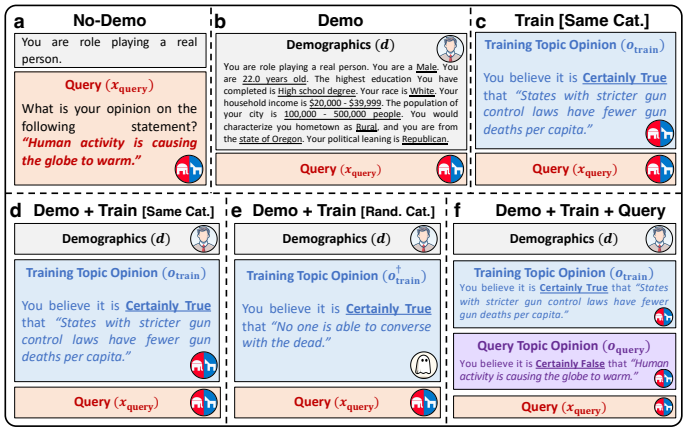

Figure 2: LLM agent construction conditions with different levels of respondent's information through in-context learning. Refer to Section 2 for detailed definitions.

2 Methods and Experimental Settings

Preliminaries: LLM Agents as Human Digital Twins We aim to construct an LLM agent i' as the i-th human's "digital twin", such that their opinions o on various topics x are aligned. We first use information about human i (e.g., their demographic information d) to create the corresponding LLM agent i', and then query the agent's opinion ($o_{i'}$) on a wide range of topics. We then evaluate the human-LLM alignment by measuring the discrepancy between the actual human opinion o_i and the LLM agent's opinion $o_{i'}$. Note that we use the term LLM-based "*agent*" to refer to the digital twin because they are designed to produce a wide range of social behaviors that emulate the human individual they role-play [18, 21, 27].

Controversial Beliefs Survey. We used the *Controversial Beliefs Survey* developed in [11], which measures opinions across $M = 64$ diverse topics pertaining to science, politics, and conspiracy theories, and more (see Appendix B for details). $N = 564$ US-based individuals rated the truthfulness of each statement on a six-point Likert scale, with higher scores indicating agreement with the rational/consensus ground truth. Demographic data (e.g., age, gender, education) were also collected. This dataset was used to construct and evaluate the LLM agents.

Belief Network Constructed with Factor Analysis. Our objective was to find independent "belief networks"–that is, groups of topics where expressed beliefs covaried across participants within each group but were independent between groups. To this end, we used a factor analysis from a prior study [11], in which the analysis produced nine independent belief networks (Figure 1), each representing an orthogonal latent factor and the corresponding topics loading highly on it. For example, the *ghost factor* included topics related to supernatural beliefs, while the *partisan factor* grouped politically polarized topics. Further details of the factor analysis are provided in Appendix F.

LLM Agent Construction. For each human respondent i, we constructed a corresponding LLM agent i' (a "digital twin") using in-context learning (ICL). For each of the nine topic categories, we selected one "training topic" (x_{train}) based on its highest factor loading within the belief network. The agent's opinion on the remaining 55 test topics ($\mathcal{X}_{\text{test}}$) was used to evaluate its alignment with human opinions. We hypothesized that initializing the agent with a human opinion on the training topic (through ICL) would generalize to related test topics in the same belief network, but not to topics from unrelated networks. To test our hypothesis, we evaluated human-LLM belief alignment under six different initialization conditions (Figure 2): **a. Baseline: No-Demo.** Agents role-play a generic person without any specific demographic or opinion information. **b. Baseline: Demo.** Agents are initialized with demographic data (d_i) only. **c. Baseline: Train [same category].** Agents are seeded with the human opinion on the training topic ($x_{\text{train}}, o_{\text{train}}$) and are assessed on test topics within the same belief network (x_{query}). **d. Demo+Train [same category].** Agents receive both demographic information (d_i) and the opinion on the training topic ($x_{\text{train}}, o_{\text{train}}$), and are tested on topics within the same belief network. **e. Baseline: Demo+Train [random category].** Agents are initialized with demographic data and an opinion from a training topic in a randomly selected belief network other than the test topics ($x^{\dagger}_{\text{train}}$, $o^{\dagger}_{\text{train}}$). **f. Upper Bound: Demo+Train+Query.** Agents are provided with both demographic data and the actual human opinions on the test topics (x_{query}, o_{query}), serving as an upper bound on alignment. Further details on LLM agent construction are in Appendix G.

Configuration for LLM Agents. We considered the following LLMs: ChatGPT (`gpt-3.5-turbo-0125`; 16), GPT-4o mini (`gpt-4o-mini-2024-07-18`), Mistral (`Mistral-7B-Instruct-v0.2`; 13), and LLaMA 3.1 (`Llama-3.1-8B-Instruct`; 24), all with temperature of 0.7. During initialization, if present, the demographic background and the training/query topic opinions were incorporated into the model's *"system messages"*. The opinion queries (x_{query}) were fed to the agent through the model's *"user messages"*.

Evaluation Metrics. We assessed the alignment between human and LLM agents' opinions by calculating the mean absolute error (MAE_{test}) across testing topics within each topic category. This measures the average discrepancy between human (o_i) and LLM ($o_{i'}$) opinions, where lower values indicate better alignment. Formally, $\text{MAE}_{\text{test}} = \frac{1}{|\mathcal{X}_{\text{test}}|} \sum_{x \sim \mathcal{X}_{\text{test}}} |o_{i,x} - o_{i',x}|$. Additionally, we computed the *relative gain* (%) to assess the improvement provided by incorporating belief network information compared to demographic data alone. That is, relative grain (%) = $(\text{MAE}_{\text{test,Baseline: Demo.}} - \text{MAE}_{\text{test,Demo+Train [same category]}})/(\text{MAE}_{\text{test,Baseline: Demo.}} - \text{MAE}_{\text{test,Upper Bound: Demo+Train+Query}}) \times 100$ (%). Full formula details and metric ranges are provided in Appendix H.

Table 1: Mean absolute error (MAE_{test}) between human respondents and the corresponding LLM agents for each topic category across various LLM agent construction conditions. The bottom row presents the relative gain (%). The lower the MAE_{test} and higher the relative gain, the higher the human-LLM alignment. The condition of our main interest (Demo + Train [Same Cat.]) is boldfaced, which also has the best alignment. GPT-4o mini and Mistral show similar results (Table 4).

Model	Condition	Topic Categories									Average
		Ghost	Psychics	Religion	Trump	Partisan	Economic	LowIndo	Health	Conspiracy	
ChatGPT	*Baselines*										
	No-Demo	2.33	2.26	1.81	1.17	1.43	1.42	1.29	1.62	1.80	1.68
	Demo	2.58	2.28	1.87	1.23	1.41	1.51	1.21	1.66	1.51	1.70
	Train [Same Cat.]	1.48	1.46	1.80	1.18	1.36	1.48	1.23	1.60	1.76	1.48
	Demo + Train [Rand. Cat.]	2.26	1.86	1.93	1.29	1.49	1.63	1.26	1.80	1.53	1.67
	Demo + Train [Same Cat.]	**1.26**	**1.27**	**1.72**	**1.14**	**1.34**	**1.23**	**1.15**	**1.53**	**1.40**	**1.34**
	Upper Bound										
	Demo + Same Train + Query	0.41	0.48	0.30	0.63	0.28	0.09	0.82	0.30	0.46	0.42
	Relative Gain (%) ↑	60.83	56.11	9.55	15.00	6.19	19.72	15.38	9.56	10.48	22.54
LLaMA 3.1	*Baselines*										
	No-Demo	2.55	2.40	1.88	1.86	2.04	2.54	1.52	1.54	2.11	2.05
	Demo	2.36	2.42	1.85	1.50	1.45	2.33	1.47	1.50	2.35	1.91
	Train [Same Cat.]	2.21	2.28	1.82	1.44	1.63	1.86	1.48	1.63	2.77	1.90
	Demo + Train [Rand. Cat.]	2.70	2.64	2.03	1.69	1.87	2.48	1.80	1.97	2.28	2.16
	Demo + Train [Same Cat.]	**2.07**	**1.88**	**1.81**	**1.19**	**1.32**	**1.69**	**1.35**	**1.07**	**2.00**	**1.60**
	Upper Bound										
	Demo + Same Train + Query	1.76	1.04	1.42	0.96	0.56	1.47	0.72	0.96	0.65	1.06
	Relative Gain (%) ↑	48.33	39.13	9.30	57.41	14.61	74.42	16.00	79.63	21.05	39.99

3 Results

Demographic information alone does not align the LLM agent's opinion. Incorporating solely the demographic information (Demo condition) fails to align LLM agents with human respondents (Table 1,4), indicated by similarly high MAE_{test} across Demo condition and No-Demo condition.

Specifying the agent's opinion on a training topic aligns other beliefs in the same network. When the LLM is provided with a human opinion on a training topic (x_{train}, o_{train}), its expressed opinions on related topics within the same belief network become more aligned with corresponding human opinions (Table 1,4). For instance, initializing an agent to believe *"some people can communicate with the dead"* (x_{train}) leads to greater alignment on related topics like *"people can project their soul out of their body"* (x_{query}). Across nine topic categories, this inclusion reduced the mean absolute error (MAE) from 1.70 to 1.34 on average (Demo condition vs. Demo+Train [same category] condition; ChatGPT), reflecting a 22.54% relative gain. However, this effect was limited to topics within the same belief network; providing a training opinion from a different category (e.g., gun control) did not improve alignment (Demo+Train [random category] condition). These results support our hypothesis that belief alignment generalizes only within related topics in the same network.

Combining demographic information and training topic opinion reaches the best alignment. While the demographic information alone is insufficient as discussed previously (Demo condition), does it offer any benefit? After removing demographic information from Demo + Train [same category] condition (i.e., Train [same category] baseline condition), MAE_{test} increases from 1.34 to 1.48 (ChatGPT), and the relative gain decreases from 22.54 % to 17.19 %. This shows that to reach the best alignment, both the training topic opinion and the demographic should be included.

Alignment does not reflect superficial repetition. We considered whether the improved alignment observed in Demo+Train [same category] condition might arise from superficial repetition of the training topic opinion (o_{train}). To test this, we conducted an experiment where the label distribution was balanced using reversed framing statements that conveyed the same semantic meaning (see Appendix J for details). This ensured the model could not align merely by repeating the provided training topic opinion. Results show that LLMs continued to demonstrate significant alignment with human opinions (Table 5), confirming that the alignment reflects the joint information $(x_{\text{train}}, o_{\text{train}})$ rather than simple repetition of the opinion label (o_{train}).

4 Conclusion

In the study, we investigated the use of empirically-derived belief networks for promoting alignment of expressed beliefs between LLM agents and twinned human participants. We showed that demographic role-playing alone does not produce significant alignment, but that initializing an agent with a human

opinion on one topic aligns opinions on related topics within the same belief network. The effect does not extend to distant topics from different belief networks. This work highlights a novel and potentially powerful means of enhancing LLM agents' alignment with human opinions.

Limitations

The scope of topics We considered just 18 topics derived from two orthogonal latent factors identified in prior work. While the Partisan topics are of public interest and the Ghost topics explore an orthogonal dimension, future research could greatly the scope of topics.

The structure of the belief network. We considered belief networks based on two highly distinct clusters to facilitate evaluation. Other studies have used more sophisticated models, such as Bayesian networks [19], which allow for precise predictions about topic interrelations. Future work could apply such methods to better characterize belief networks.

The actions of the LLM agents. Our LLM agents expressed their opinions through Likert-scale ratings. This facilitated direct comparison with human responses but may not fully capture the expression of opinions in real-world settings like social media communication. Future studies could explore more complex actions (e.g., writing social media posts) to assess their human-likeness in realistic applications.

Ethics Statement

We aim to develop LLM agents capable of simulating realistic human communicative dynamics, including the expression of potentially harmful beliefs such as misconception about the reality of global warming. Our objective is to facilitate a deeper understanding of social phenomena like misinformation spread in order to identify strategies that mitigate these challenges effectively. Note that under the current setting, the LLM agents only produce Likert-scale ratings from a fixed set of options. Therefore, they are not able to produce unexpected harmful responses. We will release our code base solely for research purposes, and adhere to the terms of use by OpenAI's API [1] and their MIT license [2], as well as Mistral AI's non-production license (MNPL) [3].

Acknowledgements

We thank the reviewers, the area chair for their feedback. This work was funded by the Multi University Research Initiative grant from the Department of Defense, W911NF2110317 (with Rogers as Co-I), Cohesive and Robust Human-Bot Cybersecurity Teams, the John S. and James L. Knight Foundation (Award Number: MSN231314), and the National Science Foundation through the Convergence Accelerator Track F: Course Correct: Precision Guidance Against Misinformation (Agency Tracking Number: 2230692; Award Number: MSN 266268).

[1] https://openai.com/policies/terms-of-use

[2] https://github.com/openai/openai-openapi/blob/master/LICENSE

[3] https://mistral.ai/licenses/MNPL-0.1.md

References

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

[2] Delia Baldassarri and Amir Goldberg. Neither ideologues nor agnostics: Alternative voters' belief system in an age of partisan politics. *American Journal of Sociology*, 120(1):45–95, 2014.

[3] Andrei Boutyline and Stephen Vaisey. Belief network analysis: A relational approach to understanding the structure of attitudes. *American journal of sociology*, 122(5):1371–1447, 2017.

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

[5] Raymond B Cattell. The scree test for the number of factors. *Multivariate behavioral research*, 1(2):245–276, 1966.

[6] Harrison Chase. Langchain, 10 2022. URL `https://github.com/langchain-ai/langchain`.

[7] Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy T Rogers. Simulating opinion dynamics with networks of llm-based agents. *arXiv preprint arXiv:2311.09618*, 2023.

[8] Yun-Shiuan Chuang, Nikunj Harlalka, Siddharth Suresh, Agam Goyal, Robert D Hawkins, Sijia Yang, Dhavan V Shah, Junjie Hu, and Timothy T Rogers. The wisdom of partisan crowds: Comparing collective intelligence in humans and llm-based agents. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*, 2024.

[9] Christopher J Devine. Ideological social identity: Psychological attachment to ideological in-groups as a political phenomenon and a behavioral influence. *Political Behavior*, 37:509–535, 2015.

[10] Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair nlp models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762, 2023.

[11] Vincent V Frigo. *An Examination of Non-Normative Belief Updating Behavior in Humans (Why Is It so Hard to Change Minds?)*. The University of Wisconsin-Madison, 2022.

[12] Caitlin E Jewitt and Paul Goren. Ideological structure and consistency in the age of polarization. *American Politics Research*, 44(1):81–105, 2016.

[13] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[14] Henry F Kaiser. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3):187–200, 1958.

[15] David M Keating. Persuasive message effects via activated and modified belief clusters: toward a general theory. *Human Communication Research*, page hqad035, 2023.

[16] OpenAI. Introducing ChatGPT. `https://openai.com/blog/chatgpt`, 2022. [Accessed 13-10-2023].

[17] Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pages 1–18, 2022.

[18] Joon Sung Park, Joseph C O'Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*, 2023.

[19] Derek Powell, Kara Weisman, and Ellen M Markman. Modeling and leveraging intuitive theories to improve vaccine attitudes. *Journal of Experimental Psychology: General*, 152(5): 1379, 2023.

[20] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR, 2023.

[21] Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. Character-llm: A trainable agent for role-playing. *arXiv preprint arXiv:2310.10158*, 2023.

[22] Seungjong Sun, Eungu Lee, Dongyan Nan, Xiangying Zhao, Wonbyung Lee, Bernard J Jansen, and Jang Hyun Kim. Random silicon sampling: Simulating human sub-population opinion using a large language model based on group-level demographic information. *arXiv preprint arXiv:2402.18144*, 2024.

[23] Amir Taubenfeld, Yaniv Dover, Roi Reichart, and Ariel Goldstein. Systematic biases in llm simulations of debates. *arXiv preprint arXiv:2402.04049*, 2024.

[24] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[25] Felicity M Turner-Zwinkels and Mark J Brandt. Belief system networks can be used to predict where to expect dynamic constraint. *Journal of Experimental Social Psychology*, 100:104279, 2022.

[26] Madalina Vlasceanu, Ari M Dyckovsky, and Alin Coman. A network approach to investigate the dynamics of individual and collective beliefs: Advances and applications of the bending model. *Perspectives on Psychological Science*, 19(2):444–453, 2024.

[27] Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, et al. Sotopia: Interactive evaluation for social intelligence in language agents. *arXiv preprint arXiv:2310.11667*, 2023.

A Related Work

Aligning human and LLM opinions. Recent studies highlight both the potential and the limitations of using LLMs to emulate human opinions [1, 20, 22, 10, 7, 8]. Argyle et al. [1] showed that LLMs conditioned on demographic backstories can emulate human voting preferences and language use, but did not investigate topic-specific opinions. Santurkar et al. [20] found that different models have different inherent opinions that often align with liberal, high-income, well-educated demographics, and that these opinions could not be shifted by providing demographic role-playing information. The current paper replicates this finding, but additionally suggests that alignment may be shifted via belief networks. To the best of our knowledge no prior work has studied such effects.

Belief networks. A great deal of prior work has studied human belief networks [3, 26, 15, 25, 19, 9, 12, 2, 4] and has developed a range of approaches beyond factor analysis for characterizing these including partial correlation networks [25] or Bayesian networks [19]. Such networks have been shown to predict "spillover effects" of attitude changes across related topics [25, 19] in human participants, where a change in a given topic can ripple through the belief network and influence related topics. In the present study, we investigate whether we can leverage the belief network derived from human data to construct LLM agents that more accurately reflect human opinions.

B Details of the Controversial Beliefs Survey.

The specific opinions we assessed were taken from the *Controversial Beliefs Survey* developed in [11]. The survey measures the direction and strength of belief across 64 topics spanning broad aspects of human knowledge, including history, science, health, religion, the supernatural, economics, politics, and conspiracy theories (see Table 2 for the full list of topics). Topics were selected to elicit a diverse range of opinions about their truthfulness (hence "controversial beliefs"). Each belief is stated as a factual proposition (e.g., "States with stricter gun control laws have fewer gun deaths per capita"), and participants rate their view about the truth of the statement on a six-point Likert scale ranging from "Certainly false" to "Certainly true." Responses with high numbers indicate agreement with the rational/consensus ground truth. The dataset also contains extensive demographic data from respondents, including age, gender, education level, household income, urban versus rural living environment, state of residence, and political leaning.

The dataset includes ratings for $N = 564$ individuals living in the US, collected from Amazon Mechanical Turk in 2018.[4]. Formally, we denote the set of 64 topics as $\mathcal{X} = \{x_j\}_{j=1}^{M}$ ($M = 64$). The survey dataset $\mathcal{D} = \{(d_i, x, o_i)|x \in \mathcal{X}\}_{i=1}^{N}$ consists of the opinion responses from N individuals, where the i-th individual having the demographic information d_i expresses an opinion o_i to the topic x.

The respondents provide their opinions ($-3 \leq o_i \leq 3, o_i \neq 0$) for each statement on a 6-point Likert scale with the values -3: Certainly false, -2: Probably false, -1: Lean false, $+1$: Lean true, $+2$: Probably true, $+3$: Certainly true. No neutral value was provided so participants must minimally lean in one direction or the other. The demographic and opinion data together were used to construct and evaluate the LLM agents (Section 2). The survey dataset can be obtained by contacting its authors [11].

C List of the 64 Topics in the Belief Survey

Table 2 shows the full statements of the 64 topics in the Belief Survey, including the topic category to which they belong according to the factor analysis result, along with whether they belong to the training or the test partition.

Topic Category	Topic Name	Topic Statement
Ghost	Dead Talk	No one is able to converse with the dead.
	Ghost	After someone has died it is not possible to see his or her ghost.
	Alien Visit	Intelligent beings from outer space have not visited the Earth via spaceships.
	Soul Walk	It is not possible for anyone to project their soul out of their body.

[4] `https://mturk.com/`

	See Future	No one is capable of having visions that accurately predict future events.
	Astrology	The position of the planets at the time of your birth has no influence on your personality.
	Roswell	No alien spacecraft has ever crashed near Roswell, New Mexico.
	Past Life	Nobody can accurately remember living a past life.
	The Secret	Strongly visualizing your fondest wish does not make it more likely to become a reality.
	Aura	Health cannot be improved by manipulating a person's aura or electrical field.
	Luck	"Lucky streaks" where random events are more likely to favor a person are not real.
	Dousing	Nobody can sense water using only a forked stick.
Psychics	Pyrokinesis	Nobody can start fires just by thinking about it.
	Thought Control	Nobody can control another's actions with their mind.
	Food	Food dropped on the ground for less than five seconds can become contaminated.
	Palm Reading	It is not possible to predict future life events from markings on a person's palm.
	Telekinesis	No one is capable of moving objects with his or her mind.
	Witches	Witches cannot influence events by using magic.
	Mind Reading	No one is capable of reading another person's thoughts.
	Moon Landing	US astronauts have landed on the moon.
	Crystals	Crystals do not have unexplained powers.
	Lightening	Lightning can strike twice in the same place.
	Alien Abd	Human beings have not been abducted by aliens from outer space.
Religion	God	God does not exist.
	Prayer	Prayer cannot cure illness.
	Angels	Angels are not real.
	Religion Explain	Religion does not provide the most accurate explanation for how the universe came into existence.
	Evil Spirit	It is not possible for a person's actions to be controlled by an evil spirit.
	Science Expl	Everything that happens can eventually be explained by science.
	Miracles	Miracles that defy the laws of nature cannot happen.
	Evolution	Species living on the Earth today have not always existed in their present form.
Trump	Homicide	In the US, about 80% of white homicide victims are killed by white people.
	Trump Inaug	More people attended the inauguration of Barack Obama than the inauguration of Donald Trump.
	Kenya	Barack Obama was born in Hawaii.
	US Employment	The US unemployment rate in 2016 was lower than 40%.
	Gov Reg	Government regulations do not always stifle economic growth.
	Holocaust	The Nazi government in Germany murdered approximately 6 million Jewish people during the second world war.
	Trump Votes	Hilary Clinton received the most overall votes in the 2016 Presidential election.
	Abortion	Strongly Republican states have higher rates of abortion than strongly Democratic states.
	Dem Guns	The official platform of the Democratic Party does not seek to repeal the 2nd Amendment.
	Health Insurance	Since the Affordable Care Act (Obamacare) passed, more Americans have health insurance.
Partisan	Gun Control	States with stricter gun control laws have fewer gun deaths per capita.
	US Deficit	The US deficit decreased after President Obama was elected.
	Globe Human	Human activity is causing the globe to warm.
	Globe Warm	The global climate is rapidly growing warmer.
	Unions	States with strong union protections have lower unemployment than states without such protections.
	Death Penalty	States that have the death penalty have higher rates of violent crime on average.
Economic	US Tax	The United States doesn't have the highest federal income tax rate of any Western country.
	Deport	President G. W. Bush deported fewer undocumented immigrants than President Obama.
	Lower Tax	Lowering taxes does not always lead to economic growth.
	Bailout	The rescue of big banks by the federal government aided recovery from the 2008 recession.
	Gold Stand	Returning to the Gold Standard would make the US more vulnerable to a recession.

LowInfo	Refugee	In 2016 fewer than 100,000 refugees from the Middle East were granted permission to live in the United States.
	US Crime	The violent crime rate in the US has declined over the past 10 years.
	Earth Age	The Earth is not around 6,000 years old.
	Human Trex	The Tyrannosaurus Rex and humans did not live on the Earth at the same time.
	Pub Priv	For a given level of education, private-sector workers typically earn more than government workers.
Health	Body Cleanse	A "body cleanse" in which you consume only particular kinds of nutrients over 1-3 days does not help your body to eliminate toxins.
	Organic	Organic foods are not healthier to eat than non-organic foods.
	Fasting	Regular fasting will not improve your health.
Conspiracy	Twin Towers	The twin towers were not brought down from the inside by explosives during the 9/11 attack.
	JFK	Only one gunman was involved in the assassination of John F. Kennedy.
	Pearl Harbor	President Roosevelt did not know about the attack on Pearl Harbor ahead of time.
	Vaccination	Vaccinations cannot cause Autism.

Table 2: The statements of the 64 topics in the Belief Survey, including the topic category to which they belong according to the factor analysis result.

D The Prompts for LLM Agent Construction Through In-context Learning

Table 3 shows the prompts we use to construct and query the LLM agents in the in-context learning setting (Section 2). Different LLM agent construction conditions include various sets of the prompt types. The parts enclosed in curly brackets "{}" are the placeholders (e.g., {demo_age}, {query_topic_statement}), where they are filled with actual information from either the respondents or the belief survey. As shown in Figure 2 and Section 2, in the **No-Demo** condition, only the "Query" prompt is included. In the **Demo** condition, both the prompt types "Demographics" and "Query" are included. In the **Demo + Train** conditions (both [same category] and [random category]), the prompt types include "Demographics", "Training Topic Opinion", and "Query". In the **Demo + Train + Query** condition, the prompt types include "Demographics", "Training Topic Opinion", "Query Topic Opinion", and "Query".

E The Choice of Number of Factors in Factor Analysis

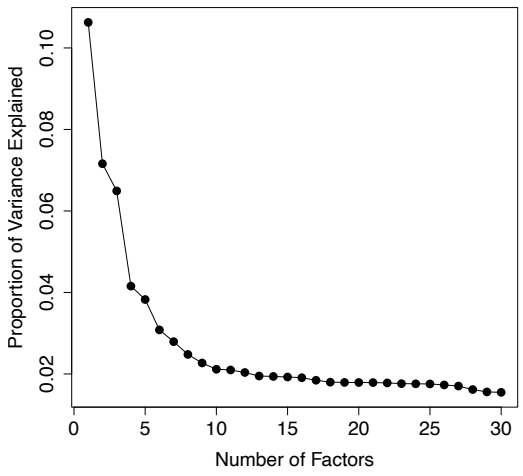

Figure 3: The scree plot of the factor analysis solution.

To determine the number of factors to retain in our factor analysis (FA), we visualize the scree plot in Figure 3. We see that the explained variance plateaus after including 9 factors (the "elbow point"). Therefore, we decide to retain 9 factors.

Table 3: The prompts used for the LLM agent construction and querying in the in-context learning setting.

Prompt Type	Message Type (LangChain)	Prompt Template	Example
Demographics	*System Message*	You are role playing a real person. You are a {demo_gender}. You are {demo_age} years old. The highest education You have completed is {demo_education}. Your race is {demo_race}. Your household income is {demo_income}. The population of your city is {demo_city_pop}. You would characterize your hometown as {demo_urban_rural}, and you are from the state of {demo_state}. Your political leaning is {demo_party}.	You are role playing a real person. You are a {Male}. You are {41} years old. The highest education You have completed is {Some college but no degree}. Your race is {White}. Your household income is $\{40,000-59{,}999\}$. The population of your city is {100,000 - 500,000}. You would characterize your hometown as {Urban (City)}, and you are from the state of {Florida}. Your political leaning is {Democrat}.
Training Topic Opinion	*System Message*	You believe that {training_topic_statement (x_{train})} is {opinion_response (o_{train})}.	You believe that {States with stricter gun control laws have fewer gun deaths per capita.} is {Probably True}.
Query Topic Opinion	*System Message*	You believe that that {query_topic_statement (x_{query})} is {opinion_response (o_{query})}.	You believe that {The global climate is rapidly growing warmer.} is {Certainly True}.
Query	*User Message*	Now, what is your opinion on the following statement using the following scale of responses? {query_topic_statement (x_{query})} is Certainly False, {query_topic_statement (x_{query})} is Probably False, {query_topic_statement (x_{query})} is Lean False, {query_topic_statement (x_{query})} is Lean True, {query_topic_statement (x_{query})} is Probably True, {query_topic_statement (x_{query})} is Certainly True. Statement: {query_topic_statement (x_{query})} Your opinion on the scale of responses:	Now, what is your opinion on the following statement using the following scale of responses? {The global climate is rapidly growing warmer.} is Certainly False, {The global climate is rapidly growing warmer.} is Probably False, {The global climate is rapidly growing warmer.} is Lean False, {The global climate is rapidly growing warmer., Probably True that {The global climate is rapidly growing warmer.} is Lean True, {The global climate is rapidly growing warmer.} is Certainly True Statement: {The global climate is rapidly growing warmer.} Your opinion on the scale of responses:

F Details of Factor Analysis Methods and Results

In the factor analysis that we leveraged from the prior study [11], they first computed correlations in the ratings produced across participants for each pair of topics, then decomposed the resulting matrix into a set of orthogonal latent factors using principal component analysis (PCA) with Varimax rotation [14].

The PCA yielded a factor *loading matrix* that encodes the loading between each topic and each latent factor (Figure 1). Nine latent factors were extracted based on the factor scree plot ([5]; Appendix E), which together accounted for 72% of the variance in the correlation matrix. The belief network surrounding these nine factors are shown in Figure 1. For example, the *ghost factor* receives high loadings from 12 topics, all pertaining to supernatural or otherworldly beliefs; the *partisan factor* receives high loadings from 6 topics on highly polarized political issues. We referred to these topics as either belonging to the *ghost topic category* or *partisan topic category*, respectively. We took these 64 topics and the corresponding latent factors as the targets for our analysis of LLM alignment.

In Figure 1 in the main text, we only show the factor loading matrix of the Ghost and the Partisan factors, and the corresponding topics. Below, we discuss the full factor analysis result.

Figure 4 shows the full factor loading matrix. The red blocks highlight strong correlations among opinions within each factor, indicating that endorsing one conception in a cluster often predicts opinion in other conceptions within the same cluster. We assign the name of each factor based on its constituent topics: Ghost, Psychics, Religion, Trump, Partisan, Economic, LowInfo, Health, and Conspiracy. The 64 topics are categorized by which factor they have the highest loadings on. For instance, the topic about communication with the dead belongs to the Ghost category because it has the highest loading on the Ghost factor (Table 2 shows the full list of topics and categories).

G Details of LLM Agent Construction and Conditions.

LLM Agent Construction. For each of the nine factors, we designated the topic possessing the highest loading as the model *training topic* (x_{train}). For each digital twin (role-playing LLM agent), the corresponding human opinion on the training topic (o_{train}) was used to customize the LLM agent through in-context learning (ICL). Human opinions on the remaining 55 testing topics $\mathcal{X}_{\text{test}}$ were not

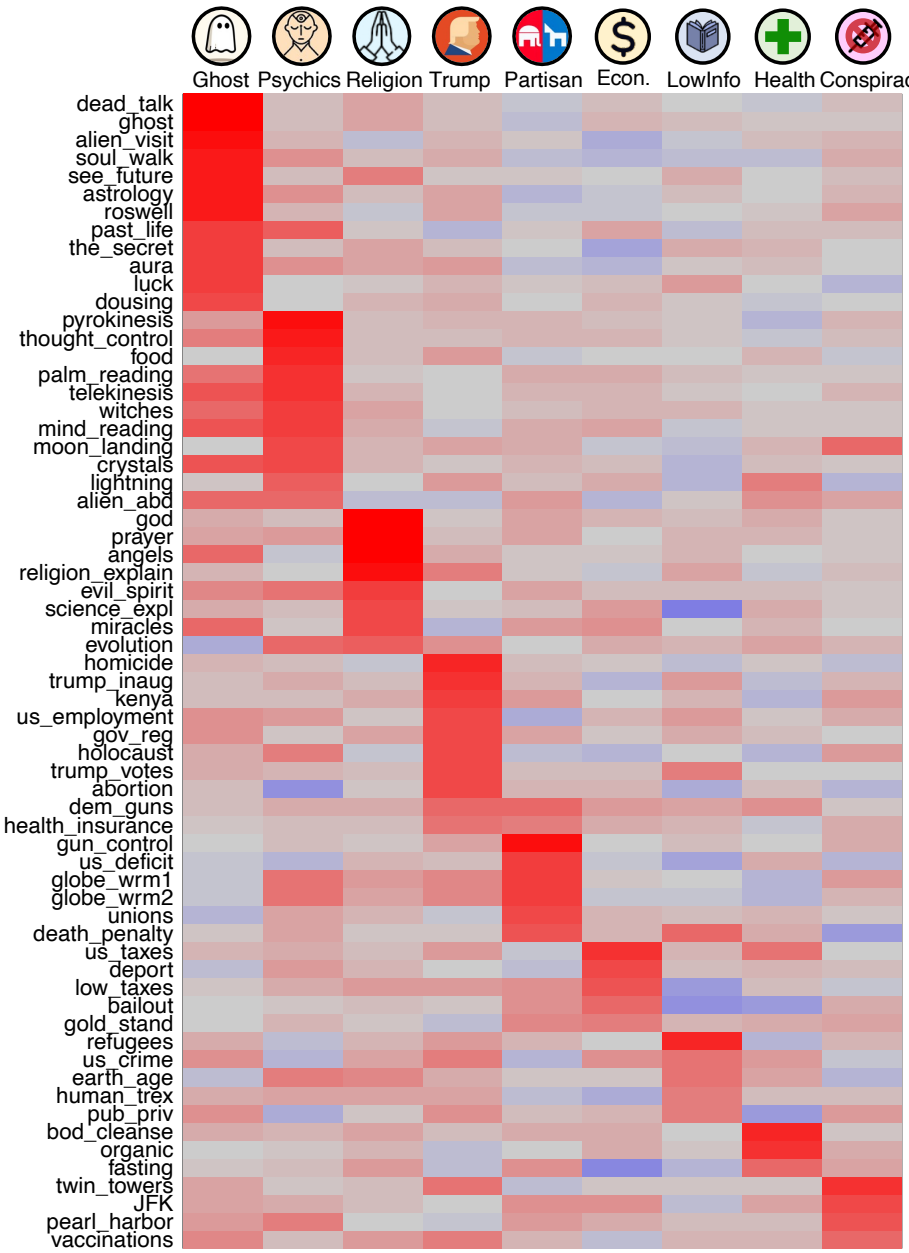

Figure 4: The factor loading matrix of the Controversial Belief Survey. The column indicates the nine factor, and the rows are the 64 topics. Red indicates topics that load highly on a factor, gray indicates near 0 loading, and blue indicates loading in the negative direction. We focus on the Ghost category and Partisan categories, highlighted by the green box and the violet box respectively. The topics in the Ghost category has minimal loading on the Partisan factor and vice versa (highlighted by the black boxes). The full statement of each topic is in Table 2 (Appendix C).

provided to the LLM agent; instead, the agent's expressed opinions o_{test} on these topics were used to evaluate their alignment with the human respondents. We hypothesized that specifying the agent's opinion on the training topic might elicit shared representation that generalizes to testing topics close within the belief network, but not those from the other belief network.

For each human respondent i, we constructed an LLM agent i' as their "digital twin," using a set of strategies described below. For each twin created under a given strategy, we queried the LLM agent for its opinions on the training and test topics (x_{query}), and measured how ratings generated by the digital twins correlate with the true opinions expressed by corresponding human respondents. We

then assessed how this measure of human-LLM belief alignment varied with different strategies for constructing the digital twin.

LLM Agent Construction Conditions. Each construction strategy involves initializing agents via ICL (Figure 2) with different information included in their *system message* (see Appendix D for the prompts). The LLM agents were constructed through LangChain [6].

a. **Baseline: No-Demo**. An LLM agent is role-playing a generic person without specific information about the human respondent (system message = '`You are role playing a real person.`"). This provides a performance floor since there is no way for the LLM to align with a corresponding human participant.
b. **Baseline: Demo.** An LLM agent is constructed to role-play the i-th respondent by adding only the demographic information (d_i) in the prompt.
c. **Baseline: Train [same category].** An LLM agent is constructed to role-play the i-th respondent by only adding the respondent's Likert-scale opinion on the training topic (x_{train}, o_{train}) and is assessed on other topics from the same topic category (x_{query}) within the belief network.
d. **Demo+Train [same category].** In addition to demographic information, the LLM receives a respondent's Likert-scale opinion on the training topic (x_{train}, o_{train}) and is assessed on other topics from the same topic category (x_{query}) within the belief network. This is the critical condition of interest.
e. **Baseline: Demo+Train [random category].** This baseline condition is similar to Demo+Train [same category], but the training topic opinion ($x^{\dagger}_{\text{train}}$, $o^{\dagger}_{\text{train}}$) belongs to a randomly selected topic category that is different from the query topic. This baseline allows us to determine whether adding respondent's Likert-scale opinion is only helpful when it belongs to the same belief network as the query topic (x_{query}).
f. **Upper Bound: Demo+Train+Query.** This condition provides the human opinion rating on both the training topic (x_{train}, o_{train}) and the query topic (x_{query}, o_{query}) during the agent construction, providing an upper bound on generalization behavior.

H Evaluation Metrics Details

Mean Absolute Error (MAE). The mean absolute error (MAE) measures the average discrepancy between the human opinions (o_i) and the LLM-generated opinions ($o_{i'}$) across all test topics ($\mathcal{X}_{\text{test}}$) in a given category. It is formally defined as:

$$\text{MAE}_{\text{test}} = \frac{1}{|\mathcal{X}_{\text{test}}|} \sum_{x \in \mathcal{X}_{\text{test}}} |o_{i,x} - o_{i',x}|$$

where $o_{i,x}$ is the human opinion on topic x, $o_{i',x}$ is the corresponding opinion generated by the LLM agent, and $\mathcal{X}_{\text{test}}$ is the set of test topics within a given belief network. The MAE value ranges from 0 (indicating perfect agreement) to 4 (indicating maximum possible disagreement).

Relative Gain (%). To assess the additional benefit of using the belief network beyond demographic information, we compute the *relative gain* (%) as the percentage improvement from the Demo. Baseline to the Upper Bound condition for the Demo+Train [same category] condition. This is calculated as:

$$\text{Relative Gain (\%)} = \frac{\text{MAE}_{\text{test,Baseline: Demo.}} - \text{MAE}_{\text{test,Demo+Train [same category]}}}{\text{MAE}_{\text{test,Baseline: Demo.}} - \text{MAE}_{\text{test,Upper Bound: Demo+Train+Query}}} \times 100\%$$

The relative gain ranges from 0% (if the belief network provides no additional benefit) to 100% (if adding the belief network achieves the same alignment as the supervised upper bound).

Table 4: Mean absolute error (MAE_{test}) between human respondents and the corresponding LLM agents for each topic category across various LLM agent construction conditions. The bottom row presents the relative gain (%). The lower the MAE_{test} and higher the relative gain, the higher the human-LLM alignment. The condition of our main interest (Demo + Train [Same Cat.]) is boldfaced, which also has the best alignment. For results with ChatGPT and LLaMA 3.1, see Table 1.

Model	Condition	Topic Categories									Average
		Ghost	Psychics	Religion	Trump	Partisan	Economic	LowIndo	Health	Conspiracy	
GPT-4o mini	*Baselines*										
	No-Demo	1.49	1.33	1.90	1.21	1.19	1.30	1.31	2.03	1.40	1.46
	Demo	1.46	1.21	1.68	1.17	1.19	1.24	1.23	1.41	1.42	1.33
	Train [Same Cat.]	1.05	0.96	1.36	1.06	1.18	1.19	1.21	1.42	1.32	1.19
	Demo + Train [Rand. Cat.]	1.44	1.23	1.53	1.28	1.24	1.22	1.19	1.58	1.41	1.35
	Demo + Train [Same Cat.]	**1.00**	**0.96**	**1.31**	**1.06**	**1.15**	**1.19**	**1.16**	**1.37**	**1.28**	**1.16**
	Upper Bound										
	Demo + Same Train + Query	0.04	0.05	0.03	0.64	0.01	0.02	0.14	0.04	0.14	0.12
	Relative Gain (%) ↑	32.39	21.55	22.42	20.75	3.39	4.10	6.42	2.92	10.94	13.88
Mistral	*Baselines*										
	No-Demo	1.75	1.63	1.64	1.33	1.20	1.07	1.49	1.30	1.44	1.43
	Demo	1.82	1.93	1.68	1.49	1.27	1.16	1.49	1.39	1.38	1.51
	Train [Same Cat.]	1.46	1.02	1.46	1.46	1.25	1.12	1.44	1.44	1.28	1.33
	Demo + Train [Rand. Cat.]	1.93	1.79	1.60	1.56	1.35	1.22	1.70	1.36	1.45	1.55
	Demo + Train [Same Cat.]	**1.36**	**1.71**	**1.41**	**1.05**	**1.25**	**1.12**	**1.12**	**1.32**	**1.27**	**1.29**
	Upper Bound										
	Demo + Same Train + Query	0.71	0.39	0.86	0.77	0.59	0.55	0.65	1.04	0.55	0.68
	Relative Gain (%) ↑	41.44	14.29	32.93	61.11	2.94	6.56	44.05	20.00	13.25	26.29

Table 5: Average MAE_{test} and average relative gain of the Demo+Train [Same Cat.] condition across the original condition ("[Original]") and the variant where we balance the label distribution ("[Balanced]"). Note that balancing the label distribution still maintains the superiority of Demo+Train [same category] condition when compared with the Demo condition.

Model	Demo + Train [Same Cat.]	
	[Original]	[Balanced]
ChatGPT		
Average MAE_{test}	1.34	1.41
Average Relative Gain (%) ↑	22.54	22.19
GPT-4-o-mini		
Average MAE_{test}	1.16	1.21
Average Relative Gain (%) ↑	13.88	9.91
Mistral		
Average MAE_{test}	1.29	1.31
Average Relative Gain (%) ↑	26.29	24.67
LLaMA 3.1		
Average MAE_{test}	1.60	1.71
Average Relative Gain (%) ↑	39.99	23.93

I Results with GPT-4o mini and Mistral

J Details of the Balanced Label Distribution Experiment

Does increased alignment following the Demo+Train [same category] condition arise from a model tendency to simply repeat the opinion providing for the training topic? Such a pattern might appear to lead to increased alignment simply because the training topic opinion, by definition, correlates with opinions on other topics in the same belief network. To address this concern, we conducted an additional experiment in which we balanced the label distribution in the context by constructing reversed framing statements that entail the same semantic meaning. We then included both the original and reversed framing statements in the context. For example, for the original statement "You believe it is *certainly true* that 'States with stricter gun control laws have *fewer* gun deaths per capita"', the reversed frame stated "You believe it is *certainly false* that 'States with stricter gun control laws have *more* gun deaths per capita"'. Both statements were included in the context in random order so the LLM cannot show increased alignment by merely repeating the training topic opinion. Table 5 shows that the LLMs continue to show significant alignment with human opinions (low MAE_{test}) in this case, an effect that must reflect the meaning of the joint information $(x_{\text{train}}, o_{\text{train}})$ rather than the opinion label o_{train} alone.