Beyond Demographics: Human-like LLM Agent Construction through Infusing Topic-Specific Opinions based on Belief Network

Anonymous ACL submission

Abstract

Creating human-like large language model (LLM) agents is crucial for faithful social simulation. Previous study has shown that when LLM agents role-play people's demographic information, their opinions are not aligned with people. This study examines the potential of integrating belief networks into LLM agent construction to enhance their ability to emulate human-like opinions. Based on a human belief survey, we estimate a belief network of 18 topics using factor analysis. Our findings reveal that, while demographic data alone fails to align LLM agent with human opinions, integrating specific topic opinions improves this alignment within related topics within the belief network. Critically, this enhanced alignment does not extend to topics outside the immediate network. These results highlight the potential of belief networks in improving LLM agent's faithful social simulations.

1 Introduction

The use of Large Language Models (LLMs) agents to simulate human-like societies has gained significant interest (Park et al., 2023, 2022; Chuang et al., 2023a; Taubenfeld et al., 2024). However, when engaging in opinionated debate, role-playing LLM agents can't faithfully role-play the personas they are given (Chuang et al., 2023a; Taubenfeld et al., 2024). For example, when a group of LLM agents are debating, regardless of the personas, they tend to endorse the model's inherent biases (e.g., denying global warming conspiracy). The model inherent biases may be introduced to ensure the safe usage of the model, but they limit the human-likeness of the LLM agents (Feng et al., 2023).

To systematically devise methods to increase the opinion alignment between role-playing LLM agents and human, we first need a benchmark dataset to quantify such alignment. Human opinion surveys are ideal benchmarks because they include both human respondents' actual demographic information and their opinions (Santurkar et al., 2023; Sun et al., 2024). Santurkar et al. (2023) show that different LLM models inherently reflect the biases of specific demographic groups, with human-feedback-tuned models reflecting opinions from liberal and well-educated demographics. Critically, they show that having LLMs to role-play the demographic data is insufficient in aligning their opinions with different human populations. Similarly, Sun et al. (2024) evaluate LLM agents' opinions across ten controversial topics when role-playing human respondents' demographic information. They show that LLM agents only accurately reflect human opinions on one out of ten topics. Overall, the LLM agents tend to adopt progressive stances (e.g., race diversity, gender equality) regardless of the demographic background they role-play.

If demographic information alone is insufficient for constructing LLM agents with human-like opinions, what additional contextual information is required? The study of "belief network" in human cognition research suggests that beliefs are not independent but are interconnected within a network that imposes constraints on which opinions are typically held together. (Boutyline and Vaisey, 2017; Vlasceanu et al., 2024; Keating, 2023; Turner-Zwinkels and Brandt, 2022). For instance, beliefs supporting increased social welfare often correlate with preferences for larger government roles. Inspired by the work on belief network, we explore the potential of incorporating specific topic opinions into LLM agent construction, with the aim to allow their opinions to extend to other interconnected topics within the belief network.

To our best knowledge, our study is the first to leverage the potential of belief network in constructing human-like LLM agents. To this end, we construct a simple belief network using factor analysis (FA). Factor analysis allows us to identify the relationships between beliefs and their underlying latent factors. FA reveals two orthogonal factors,

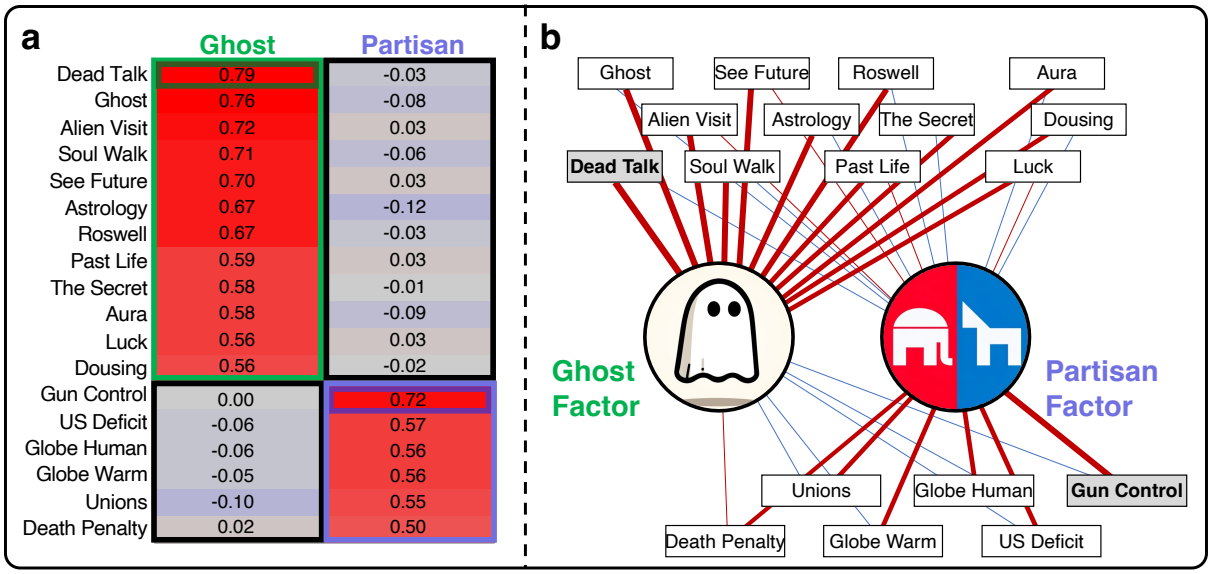

Figure 1: The belief network estimated by factor analysis from human respondents' responses on the Belief Survey. (b) Partial factor loading matrix that includes the columns for these Ghost (**green**) and the Partisan (**violet**) factors and the rows for topics that belong to these two factor categories (i.e., having the highest loadings on one of these two factors). Red indicates topics that load positively on a factor, gray indicates near 0 loading, and blue indicates loading in the negative direction. The topics in the Ghost category has minimal loading on the Partisan factor and vice versa (highlighted by the black boxes). The training topics are further highlighted by **dark green** and **purple** boxes, respectively. The full statement of the each topic is detailed in Table 2 (§A). The full factor loading matrix is in Figure 4 (§D). (a) The graphical respresentation of the belief network, where the nodes are the two latent factors, and the rectangles are the individual topics. Red and **blue** edges indicate positive and negative loadings, respectively. The width of each edge encodes the strength of the loading. The training topics are highlighted with grey backgrounds.

"Ghost" (e.g., talking to the dead) and "Partisan" (e.g., gun control), in which we explore the belief networks surrounding these two latent factors. We show that when LLM agents rely solely on demographic information, they fail to reflect human opinions. Incorporating topic-specific opinions enhances LLM alignment with human opinions on related topics in the belief network. However, this enhancement does not extend to topics that are weakly related within the network. In addition, we identify a gap relative to the supervised baseline condition, highlighting direction for improvement in future studies. This study is an initial step towards developing a more sophisticated belief network model for human-like LLM agents for social simulation.

2 Methods

2.1 Belief Survey

The Belief Survey captures the opinions of 564 individuals collected from the crowdsourcing platform Prolific [1]. The survey covers a wide range of topics, including science, politics, history, health, supernatural beliefs, economics, and religion etc. The respondents are made up of people living in the US with diverse demographic backgrounds. This survey constitutes the dataset $\mathcal{D}$, comprising tuples (d, x, o), where d represents the demographic information of each respondent, x the topic, and o the opinion expressed by the respondent on that topic. The demographic information (d) includes age, gender, education level, household income, urban versus rural living environment, state of residence, and political leaning. The 64 topics (x) are selected based on their capacity to elicit divergent opinions among the general public, with statements such as "Human activity is causing the globe to warm" and "No one is able to converse with the dead.". The full list of topics is in Table 2 (§A). The respondents provide their opinions ($-3 \leq o \leq 3$) for each statement on a 6-point Likert scale, from

[1] https://www.prolific.com/

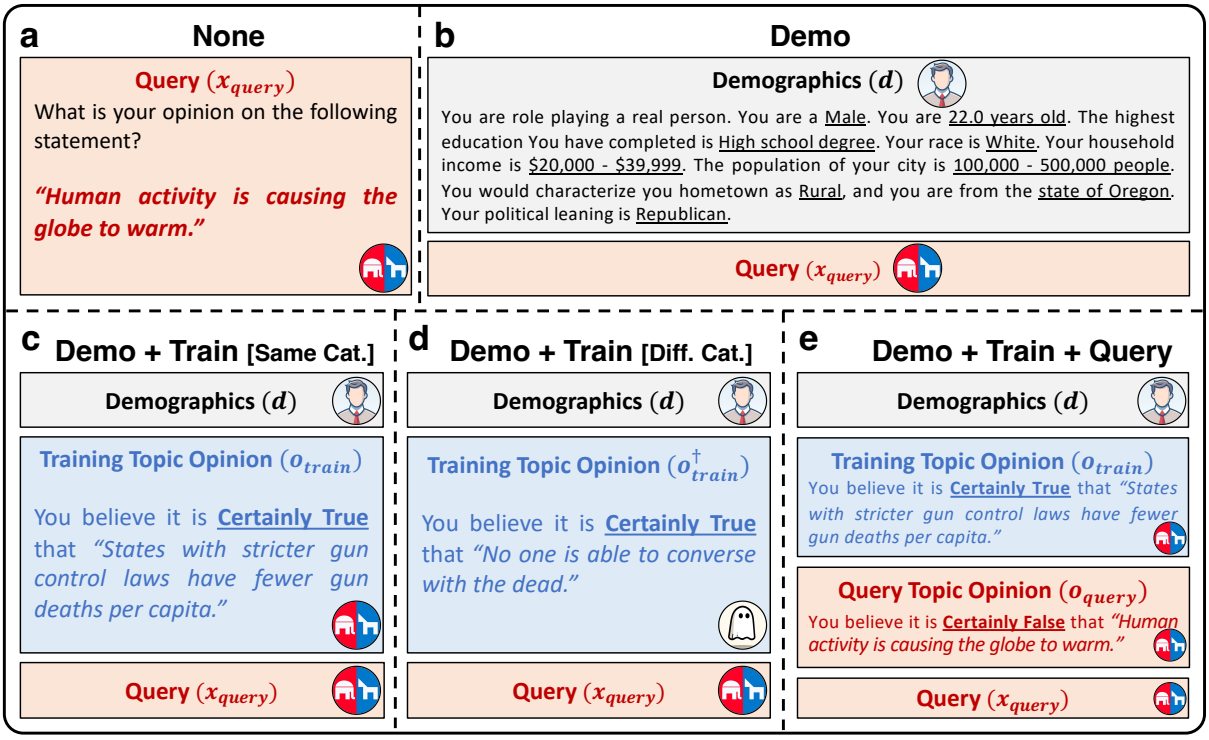

Figure 2: LLM agent construction conditions with different levels of respondent's information. (a) "None" condition without role-playing, and we directly query the LLM about its opinion on the query topic (x_{query}). (b) "Demo" with demographic information (d). (c) "Demo+Train [same category]" with demographic information, along with training topic opinion from the same topic category as the query topic (o_{train}; in this example, they both belong to the "Partisan" category). (d) "Demo+Train [different category]" with demographic information, along with and training topic opinion from a different topic category ($o^{\dagger}_{train}$); in this example, the training topic is from the "Ghost" category) (e) "Demo+Train+Query" as a supervised baseline with both training topic opinion (from the same category) and the query topic opinion (o_{query}).

-3 (Certainly false) to $+3$ (Certainly true). [2] The demographic and opinion data are then used to construct and evaluate LLM agents (§2.3).

2.2 Construct Belief Network using Factor Analysis on Belief Survey

Our research objective is to leverage belief network to create LLM agents that have human-like beliefs. To this end, we first construct a belief network based on human responses on the survey. We perform a factor analysis (FA) to uncover latent factors that explain the correlation between opinions on different topics. This yields a belief network that depicts the relationships between topics (Figure 1). We use the "varimax" variant of FA (Kaiser, 1958), which begins with Principal Component Analysis (PCA) and is followed by a varimax rotation to achieve a "simple structure" within the factor loading matrix. [3] [4]

The FA yields a factor loading matrix that encodes the association between each topic and each latent factor. We categorize the 64 topics based on the latent factor that the topic has the highest loading value on, resulting in K categories of topics. **[How do you determine K? Is it a pre-defined hyperparameter?–JUNJIE]**Within each topic category, the topics are partitioned into training set (x_{train}) and test set (x_{test}). The most representative topic for each factor (i.e., with the highest factor loading) is chosen for training and the rest for testing. This partition allows us to assess the effectiveness of incorporating a respondent's opinion on the training topic (o_{train}) into an LLM agent's construction to predict the agent's alignment with human opinions on related but unseen test topics

[2] The full list of Likert scale anchors — -3: Certainly false, -2: Probably false, -1: Lean towards false, $+1$: Lean towards true, $+2$: Probably true, $+3$ - Certainly true.

[3] The simple structure ensures that each topic is primarily associated with one single latent factor, enabling clearer categorization of topics. This procedure also ensure the latent factors are orthogonal to one another.

[4] We choose the number of factors (K) based on the scree plot (§C).

(o_{test}). The hypothesis is that opinions on a topic might elicit shared representation that generalize to other topics that are close within the belief network (i.e., sharing the same latent factor). Out of the 64 topics, we focus on a subset of 18 topics that load highly on the Ghost factor and the Partisan Factor because they form a belief network with two clear separate clusters (detailed in §3.2 and visualized in Figure 1), facilitating the evaluation of opinion generalization within and between clusters. **[I'm a bit confused with the usage of the "train/test" set. The training set x_{train} is only used for learning the FA model, but not used to update the LLM's parameter, right? –JUNJIE]**

2.3 LLM Agent Construction

For each human respondent i, we construct an LLM agent i' as the "digital twin" of the respondent. We query the LLM agents about their opinions on the 18 topics (x_{query}) of interest. We evaluate the impact of integrating different components of respondent's information into the LLM agent construction with the following conditions (Figure 2). **1. Demo Condition:** LLM agents were constructed using solely the demographic information of respondents (d), e.g., age, gender, education, race, state of residence, and political leaning. **2. Demo+Train Condition [same category]:** In addition to demographic information, we incorporate respondents' Likert-scale opinions on the training topic category (o_{train}) which belongs to the same topic category as the query topic (x_{query}). **3. Demo+Train+Query Condition:** This is the supervised baseline condition where we include the Likert-scale responses on both the training topic (o_{train}) and the query topic (o_{query}) in the agent construction. This provides the agent with direct access to the respondent's opinion on the query topic. This condition serves as an upper bound to evaluate whether the agent can replicate the given opinion when prompted. In addition to the four conditions in §2.3, we additionally add two baselines to facilitate comparisons. **4. None**: The LLM, without role-playing, is directly queried about its Likert-scale opinion on the query topic. **5. Demo+Train [different category]**: Similar to **Demo+Train**, but instead of including the opinion on the training topic within the same category as the query topic, we include the training topic from another topic category ($o^{\dagger}_{train}$). This helps evaluate if the cross-topic generalizibility of opinion is bounded within adjacent topics in the belief network. The verbatim prompts are in §B.

3 Experimental Settings

3.1 Configuration for LLM Agent

LLM agents use both ChatGPT (`gpt-3.5-turbo-0125`; OpenAI, 2022) and Mistral (`Mistral-7B-Instruct-v0.2`; Jiang et al., 2023) with temperature of 0.7 . During initialization, each agent's persona (e.g., demographic background, training/query topic opinion) are incorporated into the model's *"system messages"*. The opinion queries are added to the model's *"user messages"*. The LLM agents is managed through LangChain (Chase, 2022).

3.2 Selection of Topics based on Factor Analysis Result

The factor analysis reveals nine latent factors underlying the 64 topics (the number of factors is selected based on the scree plot, see §C). 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 (e.g., talking to the dead), Psychics (e.g., mind reading etc.), Religion (e.g., God), Trump (information spread by Trump), Partisan (e.g., gun control), Economic (e.g., taxes), LowInfo (e.g., the age of the earth), Health (e.g., organic foods), and Conspiracy (e.g., vaccinations). 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.

Among the nine topic categories, we focus on the Ghost category (12 topics) and Partisan categories (6 topics), highlighted by the green box and the violet box, respectively, in Figure 1. The Ghost category is chosen because it is the most prominent factor (i.e., explains the largest variance in the survey data), and the Partisan category is selected because topics within it have minimal loading on the Ghost factor and vice versa (highlighted by the black boxes in the figure). Because the Partisan category and the Ghost category have minimal overlap, they are ideal for evaluating how well opinion generalizes within and between categories.

3.3 Evaluation Metrics

To evaluate the "human-likeness" of the LLM agents' opinions, we measure the correlation between the Likert-scale opinions of human respondents and those generated by the LLM agents. For each topic in the survey, we compute the Kendall's Tau coefficient, τ_j, between the opinions of human respondents ($o_{i,j}$) and the opinions generated by the corresponding LLM agents ($o_{i',j}$). [5]. The coefficient τ_j ranges from -1 to 1, where 1 indicates perfect agreement, -1 indicates perfect disagreement, and 0 indicates no correlation. Furthermore, to facilitate comparison across conditions, we compute a category-wise Kendall's Tau coefficient, which is derived from the opinions across all test topics within the same topic category.

4 Results

Demographic Information Alone is Insufficient in LLM Agent's Opinion Alignment As shown in Table 1, incorporating solely demographic information (the "Demo" condition) fails to align LLM agents with human respondents. In most topics, the LLM agents yield constant responses regardless of the demographic information they role-play. This lack of variability results in undefined Kendall's Tau coefficients ("NA").

Category-Specific Training Topics Enhance LLM Agent's Opinion Alignment Critically, when integrating human respondents' opinions on training topic that belong to the same topic category as the query topic ("Demo+Train [same category]" condition), the alignment between LLM agents and human opinions increases. For example, when an LLM agent is told to believe that some people can communicate with the dead (the training topic), then the LLM agent becomes likey to also believe that people can project their soul out of their body (the query topic). However, when the training topic does not belong to the same topic category as the query topic (e.g., belief about the effectiveness of gun control law; "Demo+Train [different category]" condition), the inclusion of the training topic opinion is not helpful. This supports our hypothesis – opinions on one topic encourage the LLM agents to align their opinions on topics that load on the same latent factor.

[5] Kendall's rank correlation coefficient is preferred over Spearman's rank correlation coefficient due to its robustness to ties.

Still a Gap Towards the Supervised Baseline In the supervised baseline condition where the opinion on the query topic is explicitly included (the "Demo+Train+Query" condition), we observe a significantly higher correlation between the opinions of LLM agents and human respondents across most topics. The gap between the "Demo+Train [same category]" condition and this supervised baseline highlights opportunities for future work to enhance the human-likeness of the LLM agents' opinions.

Steerability Varies by Topics Steerability—the ability of LLM agents to adapt their opinions based on personas—differs across topics. For example, on the death penalty topic, "States that have the death penalty have higher rates of violent crime on average," there is a low correlation between human respondents and LLM agents even in the supervised baseline ("Demo+Train+Query" condition). LLM agents tend to deny the statement regardless of their personas, indicating a strong inherent model bias. Future work is needed to address this limitation.

5 Related Work

Alignment of Role-playing LLM Agent Opinions with Human Recent studies highlight both the potential and the limitations of using LLMs to emulate human opinions (Argyle et al., 2023; Santurkar et al., 2023; Sun et al., 2024; Feng et al., 2023; Chuang et al., 2023a,b). Argyle et al. (2023) introduce the concepts of "algorithmic fidelity" and "silicon sampling", where LLMs, conditioned on demographic backstories, can emulate human voting preference and language use when describing different political parties. However, they do not investigate topic-specific opinions. Santurkar et al. (2023) find that different models have different inherent opinions that align with different specific demographic groups. For example, human-feedback-tuned LLMs reflect opinions predominantly from liberal, high-income, well-educated demographics. Critically, they show that providing demographic information to the LLM could not effectively steer the LLM's opinions. Notably, however, to the best of our understanding, no existing work has studied the potential of belief networks in LLM agents' opinion alignment (detailed below).

Belief Network Human beliefs are interrelated within the structure of "belief networks" (Boutyline and Vaisey, 2017; Vlasceanu et al., 2024; Keating, 2023; Turner-Zwinkels and Brandt, 2022; Powell

Category	Topic	Condition for LLM Agent Construction									
		ChatGPT (`gpt-3.5-turbo-0125`)					Mistral (`Mistral-7B-Instruct-v0.2`)				
		None	Demo	Demo+Train [Diff. Cat.]	Demo+Train [Same Cat.]	Demo+Train + Query	None	Demo	Demo+Train [Diff. Cat.]	Demo+Train [Same Cat.]	Demo+Train + Query
Ghost											
Train	Dead Talk	0.04	0.02	0.04	0.98	1.00	NA	NA	0.07	0.97	0.98
Test	Ghost	0.03	0.05	-0.07	**0.53**	0.75	NA	NA	NA	**0.59**	0.73
	Alien Visit	-0.08	-0.05	-0.04	**0.33**	0.63	NA	NA	NA	**0.37**	0.62
	Soul Walk	-0.05	0.06	-0.07	**0.40**	0.89	NA	NA	0.07	**0.53**	0.63
	See Future	-0.03	0.07	-0.03	**0.34**	0.80	NA	NA	-0.08	**0.38**	0.85
	Astrology	-0.04	0.06	-0.07	**0.28**	0.88	NA	NA	NA	**0.32**	0.71
	Roswell	-0.10	-0.07	0.03	**0.26**	0.85	NA	NA	NA	**0.21**	0.28
	Past Life	-0.02	0.01	0.09	**0.31**	0.79	NA	NA	-0.05	**0.17**	0.61
	The Secret	-0.01	0.05	0.02	**0.32**	0.66	NA	NA	NA	**0.07**	0.67
	Aura	0.03	0.02	-0.02	**0.25**	0.80	NA	NA	NA	**0.35**	0.62
	Luck	-0.04	0.08	-0.09	**0.23**	0.84	NA	NA	NA	NA	0.46
	Dousing	-0.02	0.03	0.00	**0.19**	0.71	NA	NA	0.01	**0.23**	0.58
Partisan											
Train	Gun Control	-0.04	0.25	0.30	0.98	1.00	NA	0.33	0.12	0.90	0.90
Test	Globe Warm	-0.09	0.27	0.27	**0.27**	0.94	NA	0.32	0.22	**0.38**	0.81
	Globe Human	-0.10	0.30	0.35	**0.35**	0.98	NA	0.31	0.33	**0.39**	0.73
	US Deficit	0.03	0.02	0.03	**0.16**	0.70	NA	NA	-0.02	**0.09**	0.70
	Unions	0.03	0.18	0.08	**0.18**	0.88	NA	0.06	0.04	**0.13**	0.78
	Death Penalty	-0.14	0.00	0.00	0.00	0.32	NA	NA	NA	NA	0.46

Table 1: Kendall's Tau coefficients between human respondents and corresponding LLM agents (powered by ChatGPT and Mistral) for each topic across various LLM agent construction conditions. In particular, the inclusion of within-cluster training topic opinions significantly increases the alignment between human and LLM agent responses within the same cluster. ("Diff. Cat." : Different Topic Category; "Same Cat.": Same Topic Category)

et al., 2023). The belief network is a graph in which each node is a topic that people have opinions on, and each edge encodes the relatedness between a pair of topics. In practice, the structure of a belief network can be formulated using various graph-based models, such as partial correlation networks (Turner-Zwinkels and Brandt, 2022) or Bayesian networks (Powell et al., 2023). Critically, the belief network define the *constraints* within which these opinions are typically held together. For example, support for increased social welfare and support for a larger government role, are positively correlated. Belief networks also predict "spillover effects" of attitude changes across related topics (Turner-Zwinkels and Brandt, 2022; Powell et al., 2023), 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.

6 Conclusion

This study investigate the integration of belief networks into Large Language Models (LLMs) agent construction to enhance their simulation of human-like opinions. By evaluating against human opinion surveys, we demonstrate that relying solely on demographic information is insufficient for LLMs agents to mimic the diverse human opinions. Through factor analysis, we identify a belief network that includes topics about two orthogonal latent factors – "Ghost" and "Partisan". With this belief network, we demonstrate that opinion on one topic generalize to opinions on nearby topics within the belief network, and that the generalizability does not extend to distant topics within the network. We are the first to highlight the potential of leveraging belief networks in enhancing LLM agents' alignment with human opinions.

Despite our advancements, there remains a gap compared to the supervised baseline condition, indicating areas for future improvement. In addition, some topics are challenging to steer even under the supervised baseline, potentially due to LLMs' inherent biases. To address these challenges, future work may need to explore strategies to override these inherent biases. Such adjustments could enable LLM agents to more faithfully emulate a diverse range of human opinions, which is a crucial building block towards human-like social simulation among LLM agents.

Limitations

The Scope of Topics In the present study, the scope of the topics is limited to 18 topics derived from two orthogonal latent factors revealed by factor analysis: "Partisan" and "Ghost". These topics

were selected because they form two separate clusters, facilitating the evaluation of opinion generalization within and between these clusters. While the Partisan topics are of public interest and the Ghost topics explore an orthogonal dimension, future research could enhance the robustness of our findings by expanding the scope of the topics.

The Structure of the Belief Network In the present study, we construct belief network based on two distinct clusters to facilitate evaluation. Nonetheless, previous studies have used more sophisticated models, such as Bayesian networks (Powell et al., 2023), which allow for precise predictions about topic interrelations. Future work can use more sophisticated structures to represent the belief network.

The Actions of the LLM Agents In the present study, the LLM agents only express their opinions through a Likert scale. While this facilitates direct comparison with human respondents' opinions, Likert scale response does not fully capture the opinions in real-world settings, e.g., social media communication. Future studies can explore more complex actions to assess their human-likeness in realistic applications.

Ethics Statement

We aim to develop LLM agents capable of role-playing diverse human opinions. This may include potentially harmful beliefs such as misconception about the safety of vaccination. Our objective is not to endorse such biases but to facilitate a deeper understanding of social phenomena like misinformation spread. By simulating agents with resistant viewpoints, we aim to identify strategies that mitigate these challenges effectively. In addition, we will release our code base solely for research purposes. Finally, we make sure we adhere to the terms of use by OpenAI's API [6].

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

References

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

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

Harrison Chase. 2022. Langchain.

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

Yun-Shiuan Chuang, Siddharth Suresh, Nikunj Harlalka, Agam Goyal, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy T Rogers. 2023b. Evaluating llm agent group dynamics against human group dynamics: A case study on wisdom of partisan crowds. *arXiv preprint arXiv:2311.09665*.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. 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.

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. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

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

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

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

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

Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2022. 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.

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

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

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

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

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

Madalina Vlasceanu, Ari M Dyckovsky, and Alin Coman. 2024. 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.

A List of the 64 Topics in the Belief Survey

Table 2 shows the full stetements 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.
	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.
	Lightning	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 Taxes	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.
	Low Taxes	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	Refugees	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	Bod 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.
	Vaccinations	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.

B The Prompts Used for the LLM Agent Construction and Querying

Table 2 shows the prompts we use to construct and query the LLM agents. 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 §2.3, in the **"None Condition"**, only the "Query" prompt is included. In the **"Demo Condition"**, both the prompt types "Demographics" and "Query" are included. In the **"Demo + Train Condition"** (either "[same category]" or "[different 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".

C The Choice of Number of Factors in Factor Analysis

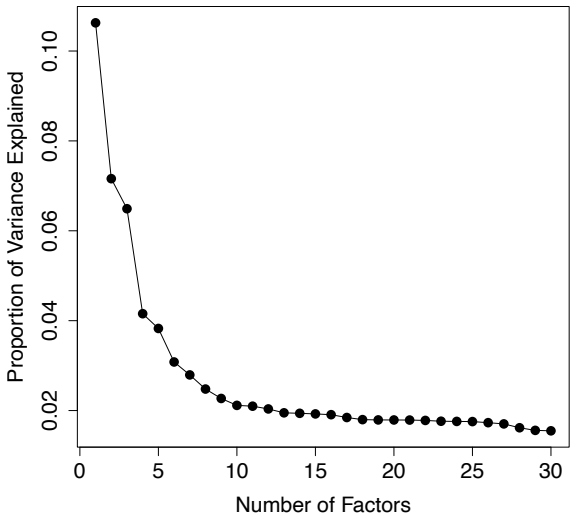

Figure 3: The scree plot of the factor analysis solution.

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.

D The Full Factor Loading Matrix

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. In Figure D, we show the full loading matrix for the 64 topics and the 9 factors.

Prompt Type	Prompt Template	Example
Demographics	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	You believe it is {opinion_response} that {training_topic_statement}.	You believe it is {Probably True} that {States with stricter gun control laws have fewer gun deaths per capita.}.
Query Topic Opinion	You believe it is {opinion_response} that {query_topic_statement}.	You believe it is {Certainly True} that {The global climate is rapidly growing warmer.}.
Query	Now, what is your opinion on the following statement using the following scale of responses? Certainly False that {query_topic_statement}, Probably False that {query_topic_statement}, Lean False that {query_topic_statement}, Lean True that {query_topic_statement}, Probably True that {query_topic_statement}, Certainly True that {query_topic_statement}. Statement: {query_topic_statement} Use the following format in providing your response: Your response: [RESPONSE SCALE OPTION]	Now, what is your opinion on the following statement using the following scale of responses? Certainly False that {The global climate is rapidly growing warmer.}, Probably False that {The global climate is rapidly growing warmer.}, Lean False that {The global climate is rapidly growing warmer.}, Lean True that {The global climate is rapidly growing warmer., Probably True that {The global climate is rapidly growing warmer.}, Certainly True that {The global climate is rapidly growing warmer.} Statement: {The global climate is rapidly growing warmer.} Use the following format in providing your response: Your response: [RESPONSE SCALE OPTION]

Table 3: The prompts used for the LLM agent construction and querying.

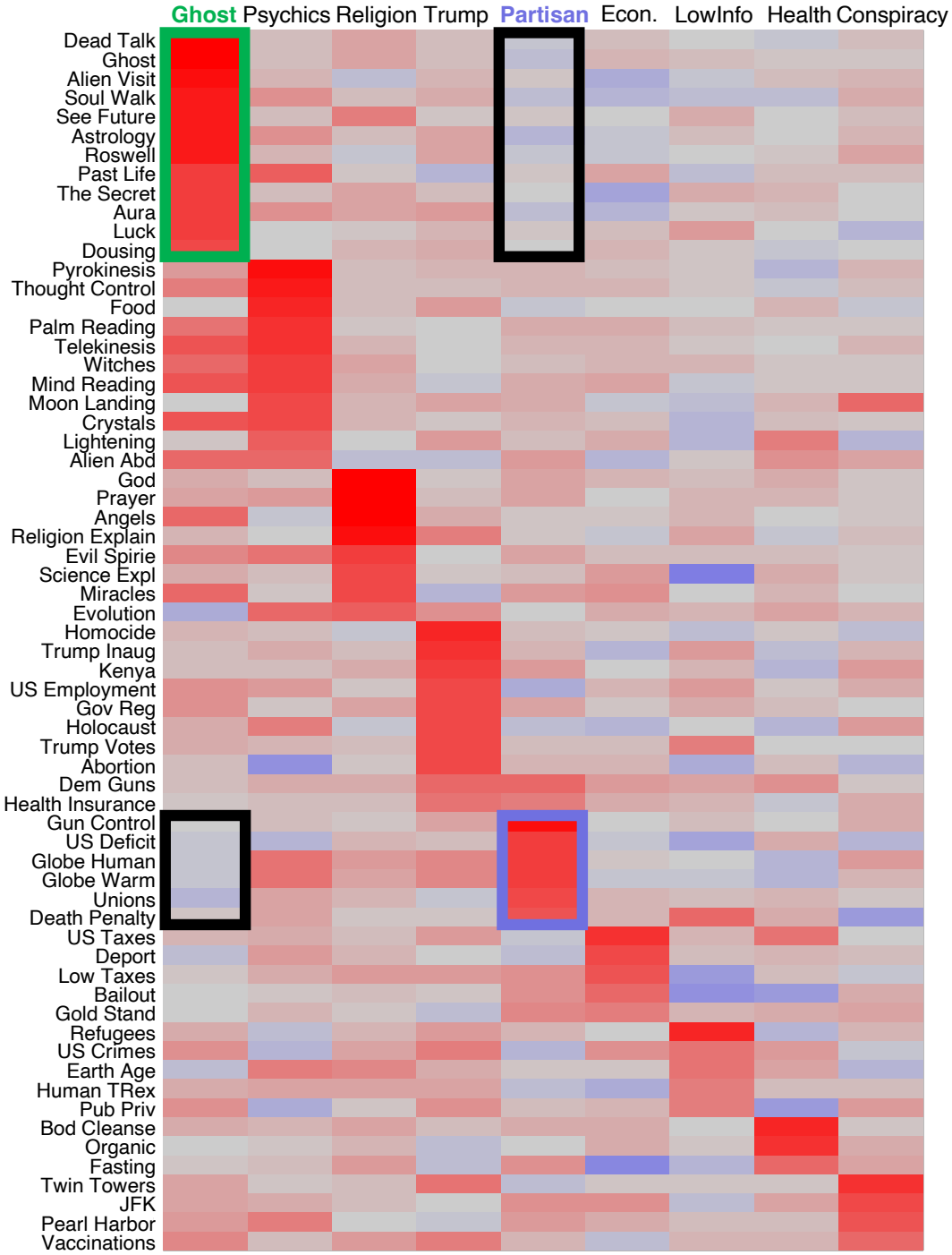

Figure 4: The factor loading matrix of the Belief Survey. The column indicates the nine factor, and the rows are the 64 topics. Red indicates topics that load highly on a factor, gray indicates near 0 loading, and blue indicates loading in the negative direction. We focus on the Ghost category and Partisan 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 (§A).