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

Anonymous ACL submission

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 18 topics loading on two 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.

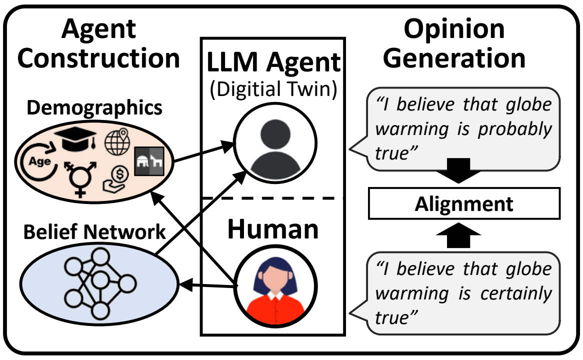

Figure 1: An LLM agent i' is constructed as the "digital twin" of a human respondent i, based on their demographic information and belief network estimated from a belief survey. We then evaluate the alignment between the opinions generated by the agent ($o_{i'}$) and those expressed by the corresponding human respondent (o_i).

1 Introduction

With rapid advances in large language models (LLMs), there has grown increasing interest in using these technologies to simulate and understand dynamics of human communication and persuasion (Park et al., 2023, 2022; Chuang et al., 2023; Taubenfeld et al., 2024). Contemporary LLMs can be prompted to role-play as individuals with particular demographic traits, sometimes then producing patterns of behavior that seem remarkably 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 (Chuang et al., 2024). 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, Santurkar et al. (2023) found that 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 does not remediate this tendency. Similarly, Sun et al. (2024) had LLMs offer opinions on controversial issues while role-playing as humans with varying demographic characteristics, and found that the model only reflected corresponding human opinions on one of the ten total topics. Chuang et al. (2023) additionally found that, even when seeded with prompts specifying an initial belief that runs contrary to social consensus (e.g., "global warming is a hoax"), LLMs quickly revert to the accepted ground-truth attitude after repeated interactions with other agents. 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 fairness and value alignment, 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 studies of human *belief networks*: the empirical observation that beliefs on different topics are not distributed at random across the population, but tend to cohere together in patterns of high-order covariation (Boutyline and Vaisey, 2017; Vlasceanu et al., 2024; Keating, 2023; Turner-Zwinkels and Brandt, 2022). 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.

To test this idea, we considered a simple belief network constructed in prior work by applying factor analysis to a dataset measuring human beliefs across a diverse array of topics (Frigo, 2022). Factor analysis decomposes patterns of covariation among expressed beliefs, identifying relationships between the beliefs themselves and a set of underlying latent factors. From this analysis we identified two orthogonal factors, each receiving high loadings from several controversial beliefs, and with no overlap between the beliefs loading highly on each. These included a *ghost factor* grouping beliefs in various supernatural phenomena (e.g., talking to the dead) and a *partisan factor* grouping beliefs that are typically politically polarizing in the US (e.g., effectiveness of gun control). We then considered how well the opinions of contemporary LLMs align with human participants when prompted (a) with no role-playing information, (b) with demographic information only, or (c) with demographic information plus a corresponding belief on a single topic that aligns strongly with either the ghost factor or the partisan factor in the belief network. When seeding each model with such a belief, we additionally compared the effects of in-context learning (i.e., prompting) versus supervised fine-tuning. The results suggest that attention to empirically-derived human belief networks may provide a useful strategy for human-LLM alignment, more so than demographic role-playing.

2 Preliminaries: LLM Agents as Human Digital Twins

As depicted in Figure 1, 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 $\text{Dist}(o_i, o_{i'})$ 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 (Park et al., 2023; Shao et al., 2023; Zhou et al., 2023).

3 Methods

3.1 Controversial Beliefs Survey

The specific opinions we assessed were taken from the *Controversial Beliefs Survey* developed in Frigo (2022). 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 4 in §A 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

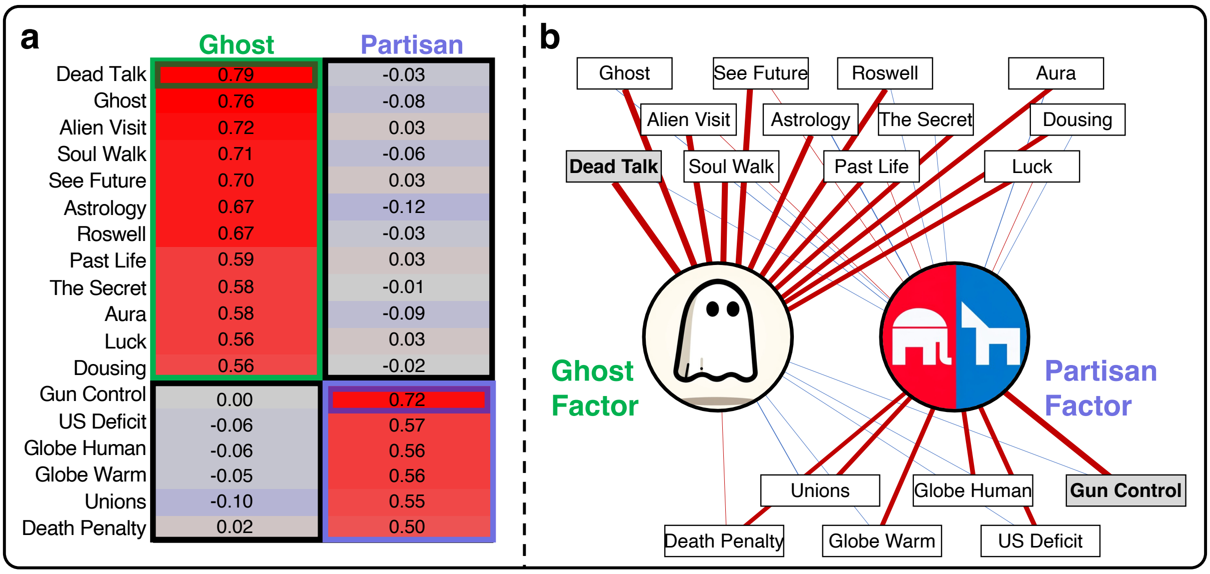

Figure 2: The belief network estimated by factor analysis from human respondents' responses on the Belief Survey. (a) Partial factor loading matrix that includes the columns for these Ghost (**green**) and the Partisan (**violet**) factors and the rows for topics that belong to these two factor categories. The full factor loading matrix is in Figure 5 (§F). 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** ("Dead Talk") and **purple** ("Gun Control") boxes, respectively. The full statement of the each topic is in Table 4 (§A). (b) The graphical respresentation of the belief network, where the central nodes are the two latent factors, and the leaves (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.

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.[1]. 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 (§3.3). The survey dataset can be obtained by contacting its authors (Frigo, 2022).

[1] https://mturk.com/

3.2 Constructing a Belief Network using Factor Analysis

Our objective was to find two independent "belief networks"–that is, two groups of topics where expressed beliefs covaried across participants within each group but were independent between groups. To this end, we relied on a previous factor analysis (Frigo, 2022) that 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 Kaiser (1958). The PCA yielded a factor *loading matrix* that encodes the loading between each topic and each latent factor. Nine latent factors were extracted based on the factor scree plot (Cattell, 1966, see §D), which together accounted for 72% of the variance in the correlation matrix. From these, we selected two factors such that topics loading highly on the first had loadings near zero on the second and vice versa. These are shown in Figure 2. The *ghost factor* receives high loadings from 12 topics, all pertaining to supernatural or otherworldly beliefs; the *partisan factor* receives high

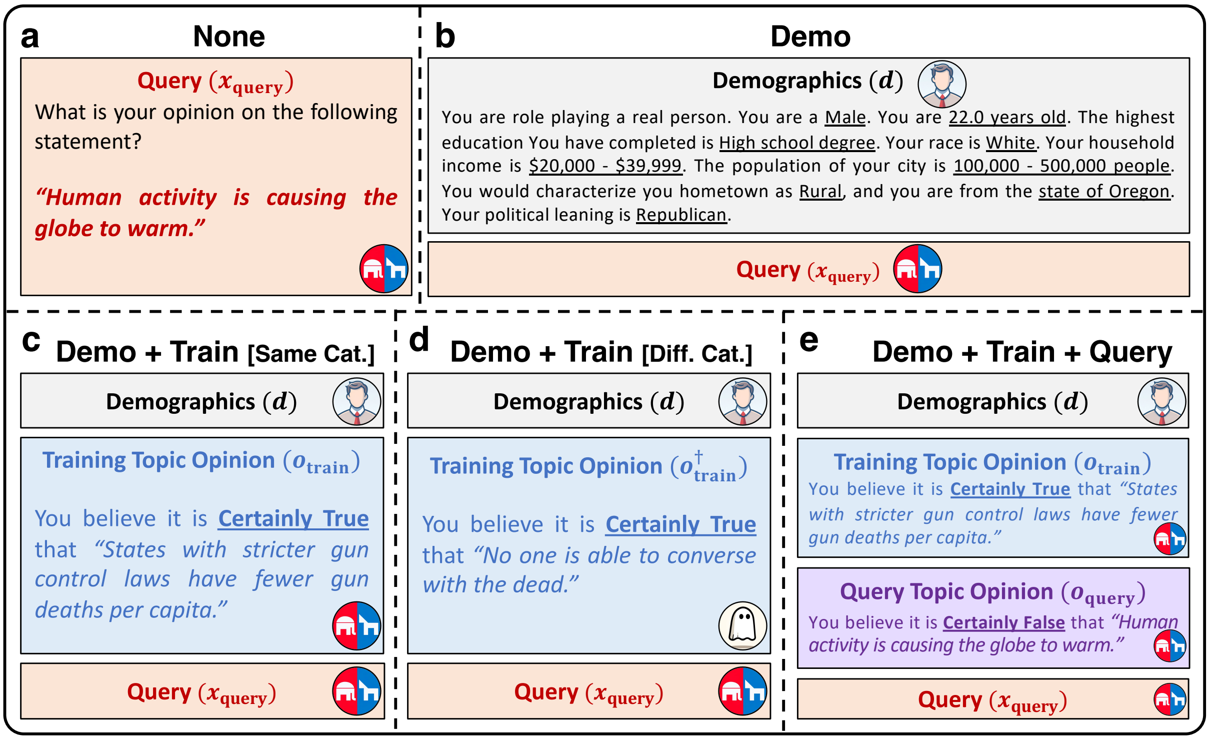

Figure 3: LLM agent construction conditions with different levels of respondent's information through in-context learning. (a) "None" condition without role-playing, and we directly query the LLM about its opinion on the query topic (x_{query}). (b) "Demo" with demographic information (d). (c) "Demo+Train [same category]" with demographic information plus training topic opinion (o_{train} on x_{train}) from the same topic category as the query topic (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 ($o^{\dagger}_{\text{train}}$ on $x^{\dagger}_{\text{train}}$) (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} on x_{query}). Everything is in the *"system message"* except the query topic, which is in the *"user message"*.

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 18 topics and the corresponding latent factors as the targets for our analysis of LLM alignment. The full factor analysis results, including the full factor loading matrix of the nine factors, can be found in §F.

3.3 LLM Agent Construction

For each factor 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 (either through in-context learning or supervised fine-tuning, see below). Human opinions on the remaining 16 testing topics x_{test} were not 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 generalize to testing topics close within the belief network (i.e., sharing the same latent factor), 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.

In-context Learning (ICL). As shown in Figure 3, these strategies involve initializing agents via in-context learning only, with different information included in their *system message* (see §4.1 and

Appendix §B for the prompts).

a. **None**: An LLM, without role-playing, is directly queried for its Likert-scale opinion on the query topic, providing a performance floor since there is no way for the LLM to align with a corresponding human participant. Note that variation may still be present due to temperature sampling (§4.1).

b. **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. **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 condition of interest.

d. **Demo+Train [different category]**: This control condition is similar to Demo+Train [same category], but assesses the LLM on topics from the opposing topic category, allowing us to determine whether the cross-topic generalization is restricted to adjacent topics in the belief network.

e. **Demo+Train+Query:** This control 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.

Supervised Fine-tuning (SFT). We also investigated whether seeding initial beliefs via supervised fine-tuning (SFT) can increase human-LLM alignment. Specifically, the correspondence between the demographic information d and the corresponding opinion o (on topic x) was used to fine-tune model weights via supervised learning, following analogous strategies to the in-context learning approaches described above. For example, for **Demo+Train [same category]**, we first construct the dataset $\mathcal{D}_{\text{SFT}} = \{(d_i, x_{\text{train},i}), o_{\text{train},i}\}_{i=1}^{N}$ for each topic category. We then fine-tuned the LLM with input context providing the demographic information along with the training topic statement (d, x_{train}), and using the corresponding human Likert-scale response o_{train} as the ground-truth output. After fine-tuning, we assessed the LLM agent's opinion on query topics x_{query} belonging to the same topic category x_{train} [2]. Likewise, for **Demo+Train [different category]**, it is similar to Demo+Train [same category] condition, but the training topic opinion ($x^{\dagger}_{\text{train}}$, $o^{\dagger}_{\text{train}}$) is from a different topic category as the query topic x_{query}. Details of the fine-tuning procedure and the corresponding prompts are in §C and §E.

4 Experimental Settings

4.1 Configuration for LLM Agents

We evaluated LLM agents using 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, the demographic background was incorporated into the model's *"system messages"*. The opinion queries (x_{query}) were fed to the agent through the model's *"user messages"*. When using in-context learning (§3.3), the training/query topic opinions were also included in the model's *"system messages"*. The LLM agents were constructed through LangChain (Chase, 2022). For our compute resources, see §G.

4.2 Supervised Fine-tuning

For LLM agents constructed through supervised fine-tuning (§3.3), we used the ChatGPT model `gpt-3.5-turbo-0125`'s fine-tuning API. Critically, because the label (i.e., opinion response o) is usually not balanced in a given topic (e.g., more people believing that ghosts are real than those who don't), we upsampled the o to ensure equal numbers of responses across the six Likert scale values. Pilot work found that, without upsampling, the fine-tuned LLM agent predominantly produced the most frequent opinion response o_{majority} in $\mathcal{D}_{\text{SFT}}$. §E lists the hyperparameters for fine-tuning.

4.3 Evaluation Metrics

To evaluate the "human-likeness" of the LLM agents' opinions, we for each topic x in the survey, we computed the Kendall's Tau coefficient, τ, between the human opinion (o_i) and that generated by the twinned LLM agent ($o_{i'}$). [3]. The coefficient τ ranges from -1 to 1, where 1 indicates perfect agreement, -1 indicates perfect dis-

[2] For example, we fine-tuned an LLM on the respondents' opinions on the training topic for the Ghost category, then queried its opinion on the test topics in the Ghost category.

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

Category	Topic	Conditions for LLM Agent Construction (In-context Learning)									
		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
	$\text{MAE}_{\text{test}} \downarrow$	[2.42]	[2.54]	[2.31]	**[1.29]**	[0.34]	[1.82]	[1.82]	[1.83]	**[1.28]**	[0.71]
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
	$\text{MAE}_{\text{test}} \downarrow$	[1.42]	[1.32]	[1.35]	**[1.25]**	[0.38]	[2.20]	[1.32]	[1.39]	**[1.28]**	[0.63]

Table 1: Kendall's τ_t between human respondents and the corresponding LLM agents (powered by ChatGPT and Mistral) for each topic across various LLM agent construction conditions through in-context learning. The bottom row presents the category-wise mean absolute error across the test topics (MAE_{test}). The higher the τ_t and the lower the MAE_{test}, the higher the human-LLM alignment. In particular, the inclusion of same-category training topic opinions significantly increases the alignment. ("Diff. Cat." : Different Category; "Same Cat.": Same Category)

agreement, and 0 indicates no correlation. To obtain a category-wise aggregated measure, we also computed the mean absolute error (MAE), $\text{MAE}_{\text{test}} = \frac{1}{|\mathcal{X}_{\text{test}}|}\sum_{x\sim\mathcal{X}_{\text{test}}} |o_{i,x} - o_{i',x}|$, which is the mean discrepancy between the opinions of human respondents and LLM agents across all test topics ($\mathcal{X}_{\text{test}}$) within the topic category.

5 Results

The results for in-context learning and supervised fine-tuning were qualitatively similar; we discuss the in-context learning results first.

Demographic information alone does not align the LLM agent's opinion. As shown in Table 1, incorporating solely the demographic information (the Demo condition) fails to align LLM agents with human respondents. The Kendall's τ are either close to zero or are undefined ("NA") due to constant responses, and the MAE_{test} of the Demo condition is also similar to the None baseline condition, indicating that the demographic information alone does not help LLM agents align with the human respondents they role-play.

Specifying the agent's opinion on a training topic aligns other beliefs in the same network. When the LLM is instructed to adopt the twinned human's opinion on the training topic (x_{train}, o_{train}) its expressed opinions on other topics in the same belief network correlate significantly (i.e., become algined) with the corresponding human opinions (Demo+Train [same category] condition; indicated by higher τ and lower MAE_{test}). For example, when an LLM agent is initialized to believe that "some people can communicate with the dead" (the training topic x_{train}), then the LLM agent becomes more likely to also believe that "people can project their soul out of their body" (the query topic x_{query}). This effect is limited to topics within the same belief network: expressed beliefs in the other topic category (e.g., about the effectiveness of gun control law; Demo+Train [different category] condition) remain uncorrelated (unaligned) with the corresponding human opinion opinion. This supports our hypothesis – opinions on one topic encourage the LLM agents to align their opinions only on topics that are adjacent in the belief network. We additionally note that such alignment is not total: human-LLM correlations in the Demo+Train [same category] condition do not reach the upper bounds established by the Demo+Train+Query control condition, highlighting opportunities for future work to further improve the alignment.

Degree of alignment reflects factor loadings. Different topics showed differing degrees of human-LLM alignment following the training-topic prompt, ranging from zero correlation for the death penalty topic ("States that have the death penalty have higher rates of violent crime on average") to a correlation of 0.53 (ChatGPT) and 0.59 (Mistral) for belief in ghosts ("After people die it is sometimes possible to see their ghost."). Yet the different topics also vary in the strength with which load on their primary factor. To assess whether this variation explains alignment patterns, we computed, across all test topics, the correlation between the topic's loading on its primary factor and its degree of alignment in the Demo+Train [same category] condition. The result showed a tight correlation between these ($r = 0.77, p < .001$), suggesting that degree of alignment following the training prompt reflects strength of the topic's participation in the corresponding belief network. This relationship does not explain all cross-topic variation; at least one topic (death penalty) showed zero alignment even when given the correct opinion in the prompt, suggesting some degree of inherent bias in model responses for certain topics.

Alignment does not reflect superficial repetition. 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 prompting contexts 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 <u>*certainly true*</u> that 'States with stricter gun control laws have <u>*fewer*</u> gun deaths per capita"', the reversed frame stated "You believe it is <u>*certainly false*</u> that 'States with stricter gun control laws have <u>*more*</u> 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 2 shows that the LLMs continue to show significant alignment with human

Category	Topic	Demo+Train condition [Same Cat.]			
		ChatGPT		Mistral	
		[Original]	[Balanced]	[Original]	[Balanced]
Ghost					
Train	Dead Talk	0.98	0.99	0.97	0.97
Test	Ghost	0.53	0.46	0.59	0.61
	Alien Visit	0.33	0.25	0.37	0.18
	Soul Walk	0.40	0.40	0.53	0.53
	See Future	0.34	0.16	0.38	0.52
	Astrology	0.28	0.13	0.32	0.32
	Roswell	0.26	0.31	0.21	0.12
	Past Life	0.31	0.32	0.17	0.18
	The Secret	0.32	0.14	0.07	0.07
	Aura	0.25	0.15	0.35	0.32
	Luck	0.23	0.03	NA	NA
	Dousing	0.19	0.24	0.23	0.32
	$\text{MAE}_{\text{test}} \downarrow$	[1.29]	[1.64]	[1.28]	[1.26]
Partisan					
Train	Gun Control	0.98	0.88	0.90	0.93
Test	Globe Warm	0.27	0.03	0.38	0.14
	Globe Human	0.35	0.12	0.39	0.21
	US Deficit	0.16	0.01	0.09	0.10
	Union Protection	0.18	0.18	0.13	0.19
	Death Penalty	0.00	0.00	NA	NA
	$\text{MAE}_{\text{test}} \downarrow$	[1.25]	[1.24]	[1.28]	[1.23]

Table 2: Kendall's τ_t between human respondents and the corresponding LLM agents (powered by ChatGPT and Mistral) for each topic across the original Demo+Train [same category] condition ("[Original]") and the variant where we balance the label distribution ("[Balanced]") in the in-context learning setting. The bottom row presents the category-wise mean absolute error across the test topics (MAE_{test}). The higher the τ_t and the lower the MAE_{test}, the higher the human-LLM alignment. Note that balancing the label distribution still maintains the superiority of Demo+Train [same category] condition when compared with the Demo condition (Table 1).

opinions (high τ and 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.

Supervised fine-tuning yields similar results. As shown in Table 3, when the agents are fine-tuned with a training topic x_{train}, they also express more human-like opinions on query topics belonging to the same belief network (i.e., higher τ and lower MAE_{test}; the Demo+Train [same category] condition), but not on those belonging to a different network (Demo+Train [different category] condition)–a pattern of results qualitatively similar to in-context learning.

6 Related Work

Aligning human and LLM opinions. 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., 2023,

Cat.	Topic	Conditions for LLM Agent Construction (SFT)			
		None	Demo	Demo+Train [Diff. Cat.]	Demo+Train [Same Cat.]
Ghost					
Train	Dead Talk	0.04	0.02	0.04	0.22
Test	Ghost	0.03	0.05	-0.08	**0.10**
	Alien Visit	-0.08	-0.05	-0.02	**0.10**
	Soul Walk	-0.05	0.06	-0.05	**0.14**
	See Future	-0.03	0.07	0.07	**0.12**
	Astrology	-0.04	0.06	0.07	**0.06**
	Roswell	-0.10	-0.07	0.05	**0.16**
	Past Life	-0.02	0.01	-0.02	**0.06**
	The Secret	-0.01	0.05	0.05	**0.14**
	Aura	0.03	0.02	-0.07	**0.06**
	Luck	-0.04	0.08	-0.06	**0.17**
	Dousing	-0.02	0.03	-0.07	**0.08**
	$\text{MAE}_{\text{test}} \downarrow$	[2.42]	[2.54]	[2.45]	**[1.65]**
Partisan					
Train	Gun Control	-0.04	0.25	0.20	0.28
Test	Globe Warm	-0.09	0.27	0.02	**0.30**
	Globe Human	-0.10	0.30	0.15	**0.30**
	US Deficit	0.03	0.02	0.01	**0.09**
	Union Protection	0.03	0.18	0.07	**0.18**
	Death Penalty	-0.14	0.00	0.00	**0.05**
	$\text{MAE}_{\text{test}} \downarrow$	[1.42]	[1.32]	[1.71]	**[1.26]**

Table 3: Kendall's τ between human respondents and the corresponding LLM agents for each topic across various LLM agent construction conditions through supervised fine-tuning. The bottom row presents the category-wise mean absolute error across the test topics (MAE_{test}). The higher the τ_t and the lower the MAE_{test}, the higher the human-LLM alignment. In particular, fine-tuning LLM with same-category training topic opinions significantly increases the alignment. The "None" and "Demo" conditions are identical to the ones in Table 1 because they are tuning-free baselines.

2024). Argyle et al. (2023) 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. (2023) 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 (Boutyline and Vaisey, 2017; Vlasceanu et al., 2024; Keating, 2023; Turner-Zwinkels and Brandt, 2022; Powell et al., 2023; Devine, 2015; Jewitt and Goren, 2016; Baldassarri and Goldberg, 2014; Brandt and Sleegers, 2021) and has developed a range of approaches beyond factor analysis for characterizing these including partial correlation networks (Turner-Zwinkels and Brandt, 2022) or Bayesian networks (Powell et al., 2023). Such networks have been shown to predict "spillover effects" of attitude changes across related topics (Turner-Zwinkels and Brandt, 2022; Powell et al., 2023) 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.

7 Conclusion

We investigated the use of empirically-derived belief networks for promoting alignment of expressed beliefs between Large Language Model (LLM) agents and twinned human participants. We showed that demographic role-playing alone does not produce significant alignment (Santurkar et al., 2023), but that initializing an agent with a human opinion on one topic then aligns opinions on nearby topics within the belief network. The effect does not extend to distant topics within the network, and varies depending the strength of the test-topic's participation in the belief network. We found similar effects for in-context learning and supervised fine-tuning, for both a proprietary and an open-source LLM. 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 (Powell et al., 2023), 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 [4] and their MIT license [5], as well as Mistral AI's non-production license (MNPL) [6].

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

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

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

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.

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

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.

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

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

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. 2023. Simulating opinion dynamics with networks of llm-based agents. *arXiv preprint arXiv:2311.09618*.

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

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

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.

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

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

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.

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

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.

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

A List of the 64 Topics in the Belief Survey

Table 4 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 4: 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 for LLM Agent Construction Through In-context Learning

Table 5 shows the prompts we use to construct and query the LLM agents in the in-context learning setting (§3.3). 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 3 and §3.3, 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** conditions (both [same category] and [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 Prompts for LLM Agent Construction Through Supervised Fine-tuning

Table 6 shows the prompts we use to construct and query the LLM agents in the supervised fine-tuning setting (§3.3). The demographic information is included in the system message in the same prompt template as in §B. For the topic-specific opinions, however, instead of including them in the prompt, we formulate them as (prompt, response) pairs for supervised fine-tuning, where prompt is the input and response is the output. The prompt templates and examples are shown in Table 6.

D The Choice of Number of Factors in Factor Analysis

To determine the number of factors to retain in our factor analysis (FA), we visualize the scree plot in Figure 4. We see that the explained variance plateaus after including 9 factors (the “elbow point”). Therefore, we decide to retain 9 factors.

E Supervised Fine-tuning Details

In this section, we elaborate the different strategies used for constructing LLM agents through supervised fine-tuning.

a. **None:** Baseline without fine-tuning, (identical to same condition in ICL.

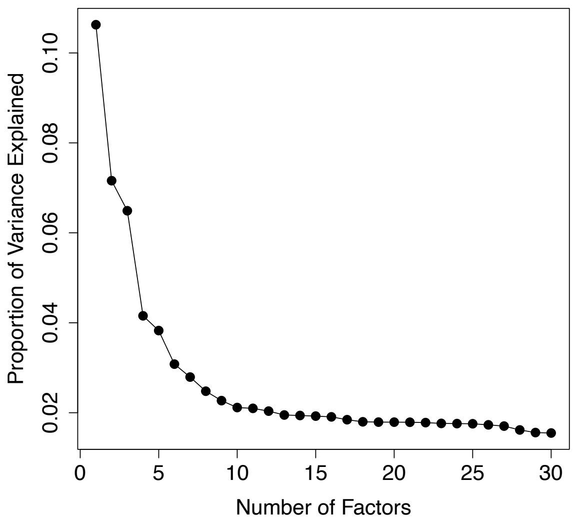

Figure 4: The scree plot of the factor analysis solution.

b. **Demo:** Baseline without fine-tuning, identical to same condition in ICL.

c. **Demo+Train [same category]:** For each topic category we constructed the dataset $\mathcal{D}_{\text{SFT}} = \{(d_i, x_{\text{train},i}), o_{\text{train},i}\}_{i=1}^{N}$. We then fine-tuned the LLM with input context providing the demographic information along with the training topic statement (d, x_{train}), and using the corresponding human Likert-scale response o_{train} as the target. After fine-tuning, we assessed the LLM agent’s opinion on query topics x_{query} belonging to the same topic category x_{train} [7]. This is the critical condition of interest that tests cross-topic generalization. The verbatim prompts are in §C.

d. **Demo+Train [different category]**: Similar to Demo+Train [same category] condition, but the training topic opinion ($x^{\dagger}_{\text{train}}$, $o^{\dagger}_{\text{train}}$) is from a different topic category as the query topic x_{query}, allowing us to assess whether generalization is restricted to topics in the same belief category.

ChatGPT (`gpt-3.5-turbo-0125`) is fine-tuned through OpenAI’s fine-tuning API [8]. These were the hyper-parameters used in fine-tuning:

- Number of Epochs: 3
- Batch Size: 1
- Learning Rate Multiplier: 2

[7]For example, we fine-tuned an LLM on the respondents’ opinions on the training topic for the Ghost category, then queried its opinion on the test topics in the Ghost category.

[8]https://platform.openai.com/docs/guides/fine-tuning

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:

Table 5: The prompts used for the LLM agent construction and querying in the in-context learning setting.

Prompt Template	Example Prompt	Response Template	Example Response
What is your opinion on the following statement using the following scale of responses? Certainly False that {query_topic_statement (x_{query})}, Probably False that {query_topic_statement (x_{query})}, Maybe False that {query_topic_statement (x_{query})}, Maybe True that {query_topic_statement (x_{query})}, Probably True that {query_topic_statement (x_{query})}, Certainly True that {query_topic_statement (x_{query})} Statement: {query_topic_statement (x_{query})}. Please choose your response from the following list of options: Certainly False, Probably False, Maybe False, Maybe True, Probably True, Certainly True.	What is your opinion on the following statement using the following scale of responses? Certainly False that {States with stricter gun control laws have fewer gun deaths per capita}, Probably False that {States with stricter gun control laws have fewer gun deaths per capita}, Maybe False that {States with stricter gun control laws have fewer gun deaths per capita}, Maybe True that {States with stricter gun control laws have fewer gun deaths per capita}, Probably True that {States with stricter gun control laws have fewer gun deaths per capita}, Certainly True that {States with stricter gun control laws have fewer gun deaths per capita} Statement: {States with stricter gun control laws have fewer gun deaths per capita} Please choose your response from the following list of options: Certainly False, Probably False, Maybe False, Maybe True, Probably True, Certainly True.	My Response: {opinion_response}	My Response: {Certainly True}

Table 6: The prompts used for the LLM agent construction and querying in the supervised fine-tuning setting.

F The Full Factor Analysis Results

In Figure 2 in the main text, we only show the factor loading matrix of the Ghost and the Partisan factors, and the corresponding topics. In this section, we discuss the full factor analysis result.

The factor analysis reveals nine latent factors underlying the 64 topics. Figure 5 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 4 shows the full list of topics and categories).

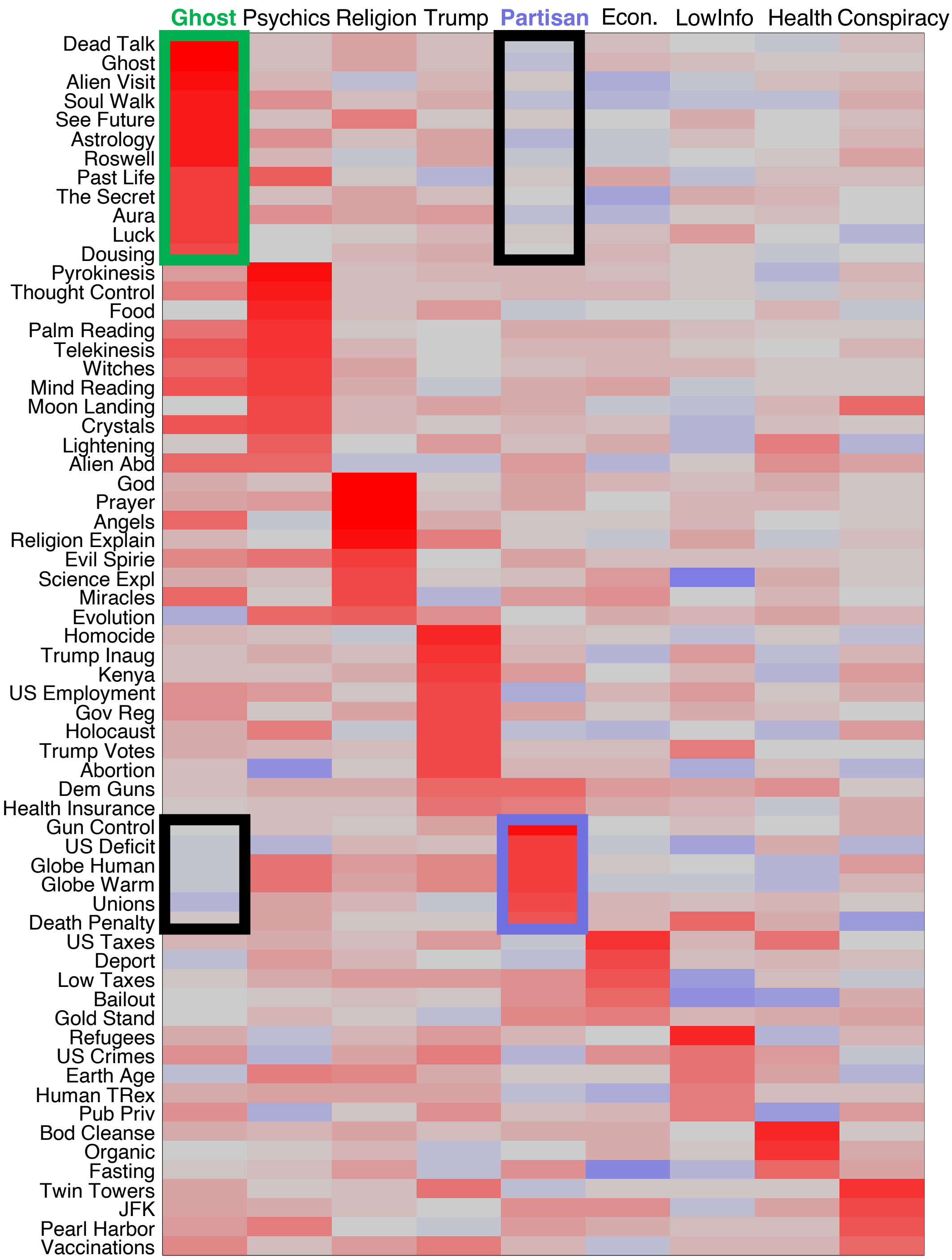

Figure 5: The factor loading matrix of the Controversial Belief Survey. The column indicates the nine factor, and the rows are the 64 topics. Red indicates topics that load highly on a factor, gray indicates near 0 loading, and blue indicates loading in the negative direction. We focus on the Ghost category 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 4 (§A).

G Compute Resources

We ran all experiments with Mistral on a GPU machine equipped with 1x NVIDIA A100. The experiments with ChatGPT cost about 300 USD.