nature human behaviour

Article

https://doi.org/10.1038/s41562-025-02228-z

A semantic embedding space based on large language models for modelling human beliefs

Received: 30 August 2024

Accepted: 25 April 2025

Published online: 04 June 2025

Check for updates

Byunghwee Lee, Rachith Aiyappa, Yong-Yeol Ahn, Haewoon Kwak ✉ & Jisun An ✉

Beliefs form the foundation of human cognition and decision-making, guiding our actions and social connections. A model encapsulating beliefs and their interrelationships is crucial for understanding their influence on our actions. However, research on belief interplay has often been limited to beliefs related to specific issues and has relied heavily on surveys. Here we propose a method to study the nuanced interplay between thousands of beliefs by leveraging online user debate data and mapping beliefs onto a neural embedding space constructed using a fine-tuned large language model. This belief space captures the interconnectedness and polarization of diverse beliefs across social issues. Our findings show that positions within this belief space predict new beliefs of individuals and estimate cognitive dissonance on the basis of the distance between existing and new beliefs. This study demonstrates how large language models, combined with collective online records of human beliefs, can offer insights into the fundamental principles that govern human belief formation.

Beliefs are foundational for human cognition and decision making. The term 'belief' refers to a conviction that something is true or exists[1], or a confidence in the rightness of something or someone[2]. Beliefs guide how individuals derive meaning, shape behaviour, filter information and form social connections that define their communities[3–6].

Research across disciplines has advanced our quantitative understanding of human beliefs. The growth of digital behavioural data and new analytical tools has enabled novel approaches to studying belief systems, ranging from mapping individual belief structures to analysing how beliefs spread through societies. Notable approaches include modelling of belief dynamics using social network diffusion models[7–12] and frameworks that integrate both individual belief systems and social influence mechanism, along with their empirical applications[3,4,13–19]. In parallel, researchers in natural language processing and social media analytics have developed methods to detect and predict individuals' beliefs through their digital footprints and textual expressions[20–26].

Studies have revealed that human beliefs are interconnected, and understanding these relationships is crucial for comprehending how beliefs form, update and propagate alongside associated behaviours. For instance, individuals sharing similar beliefs can influence each other's lifestyle choices, leading to clustered behaviours and preferences within social groups[3]. This explains why seemingly unrelated beliefs (or preferences), such as being liberal and drinking lattes, can become associated. The associative diffusion model[4] also demonstrates how relationships between beliefs shape cultural differentiation. This model suggests that cultural differences emerge primarily through the transmission of perceived compatibility between beliefs and behaviours, rather than through the direct transmission of the beliefs themselves. This process can occur even in communities lacking pre-existing social clusters. Another experiment suggests that a small number of early movers can initiate belief–ideology associations that subsequently develop into strong partisan alignments, even for beliefs fundamentally unrelated to political ideology[13]. Recent

Center for Complex Networks and Systems Research, Luddy School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN, USA. ✉e-mail: hwkwak@iu.edu; jisunan@iu.edu

https://doi.org/10.1038/s41562-025-02228-z

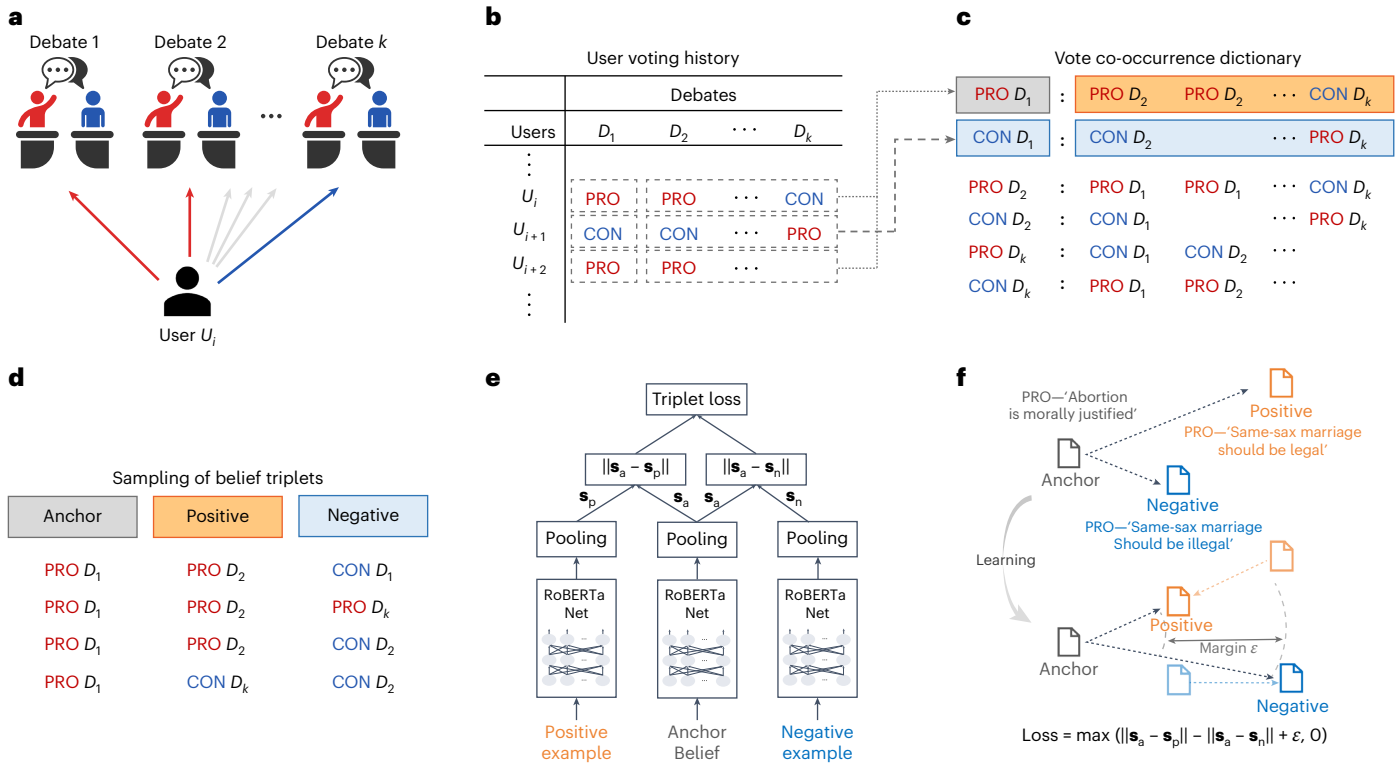

Fig. 1 | Fine-tuning S-BERT with belief triplets. a, An illustration of a user's expressed positions in multiple debates. For each debate topic, users can vote for the PRO or CON side. **b**, Voting histories of users represented as a matrix. **c**, The vote co-occurrence dictionary captures how users voted on other beliefs, given their PRO/CON vote on a certain belief. **d**, From the vote co-occurrence dictionary, belief triplets are sampled. Each belief triplet is composed of an anchor belief in addition to a positive and a negative belief in relation to the anchor. **e**, A pre-trained S-BERT model is fine-tuned with the belief triplets in the form of triplet network. **f**, An illustration of the learning process happening within S-BERT. To minimize the triplet loss, an anchor belief is drawn closer to beliefs with positive relationships.

theoretical frameworks have integrated individual belief systems with social network influences, modelling belief dynamics through the lens of dissonance theory and network imbalance[14–19]. These models illuminate how discrepancies and imbalances between beliefs shape the ultimate distribution of social beliefs. Collectively, these studies highlight that understanding the interconnected nature of beliefs is crucial for explaining societal fragmentation and polarization.

Yet, the relational landscape of human beliefs remains incompletely mapped, and our understanding of how belief interactions form is still limited. Despite the advances in belief quantification and modelling, substantial challenges still persist. A primary challenge in modelling belief systems lies in representing the nuanced relationships between beliefs. Network-based approaches to modelling human attitudes typically rely on survey data on specific issues to explore belief interrelationships (for example, through partial correlation between questionnaire responses[3,14,17]). However, survey-based methods face inherent scalability limitations when considering the entire 'space' of beliefs. Capturing relationships among vast numbers of important beliefs and incorporating new beliefs into existing systems—a process known as inductive reasoning—pose considerable challenges.

Here, we construct a robust and general representation space for beliefs that enables both continuous and inductive reasoning about beliefs and their relationships. Drawing inspiration from vector space models that encode semantic and contextual relationships between words into geometric relations[27–29], our approach leverages large language models (LLMs), combined with revealed belief trajectories extracted from online debates. This methodology creates a continuous, high-dimensional representation space. Our framework uses an empirical dataset of multiple beliefs held by individuals from an online debate forum to fine-tune a pre-trained LLM—a model that initially trained on extensive text corpora to capture broad linguistic patterns. The resulting belief-LLM translates belief statements into embedding vectors, creating a space where spatial distances reflect both semantic relationships and socially perceived relevance between beliefs. It also enables representing individuals as vectors in belief space, allowing inference of their implicit beliefs and exploration of belief adoption processes.

In this study, we aim to propose a novel framework for constructing a robust belief embedding space using LLMs integrated with online user activity data. We investigate the emergent structural characteristics in this belief embedding space, focusing on clustering and polarization patterns around social issues. Through evaluation of the embedding space's effectiveness in inferring individuals' stances on new debate topics from their existing beliefs, we address a fundamental question, namely: what mechanisms underlie human belief selection? We explore this by analysing the factors that enable accurate belief prediction.

Results

Generating and validating the belief embedding space

Our first goal is to construct a representation space that captures the interdependencies between diverse beliefs. We achieve this by fine-tuning pre-trained LLMs using contrastive learning[30]. This approach enables models to learn a representation space by attracting similar (positive) belief pairs while repelling dissimilar (negative) ones, allowing us to distinguish commonly shared belief pairs from those that are in opposition.

We leverage user participation records from an online debate forum, Debate.org (DDO)[20,21,31]. This dataset consists of online debates and corresponding voting records of the users; users can express their position by directly participating as debaters or voting for the

Table 1 | Performance of various LLMs in the belief triplet evaluation task for both the training and test sets

Model type	Pre-trained model	Triplet evaluator		GLUE-STSBSpearman correlation
		Training set	Test set	
S-BERT (fine-tuned)	roberta-base-nli-stsb-mean-tokens	0.946 (0.001)	0.674 (0.002)	0.718 (0.005)
S-BERT (before fine-tuning)	roberta-base-nli -stsb-mean-tokens	0.397 (0.001)	0.376 (0.003)	0.877
BERT (fine-tuned)	bert-base-uncased	0.933 (0.003)	0.669 (0.004)	0.476 (0.045)
BERT (before fine-tuning)	bert-base-uncased	0.376 (0.001)	0.356 (0.002)	0.615

Scores represent the average accuracy obtained from a fivefold validation task. A higher accuracy indicates that the model more accurately distinguishes positive examples from negative ones for a given anchor belief. Numbers in parentheses denote standard deviations. The last column demonstrates performance in the semantic textual similarity benchmark on general language understanding evaluation datasets (GLUE-STSB) task[35], where the goal is to estimate semantic textual similarity between two texts. Performance is assessed through Spearman correlation between the human-annotated benchmark score (rated on a scale of 1–5) and the cosine similarity between the vector representations of the two texts, as generated by LLMs.

PRO, CON or TIE position in the debates. We consider both debaters and voters simply as voters since they support a particular position in the debate. After pre-processing, we obtained a dataset of 59,986 unique debate titles voted on 197,306 times by 40,280 unique users, retaining only PRO and CON votes, which was used for fine-tuning LLMs (Methods).

We operationalize each individual's belief as their expressed agreement or disagreement with a certain debate title. We transform voting records (PRO/CON) of users into belief statements by using predefined templates. For example, if a user voted PRO (CON) to a debate titled 'Abortion is morally justified', we create a belief statement for the user as 'I agree (disagree) with the following: abortion is morally justified' (see the Supplementary Information for template variability on belief embeddings).

To encode these belief statements, we employ a pre-trained Sentence-BERT (S-BERT) model with RoBERTa[32,33]. Unlike the original BERT model[34], which is focused on token-level tasks, S-BERT is designed for generating semantically meaningful sentence-level embeddings and allows for efficient fine-tuning using sentence-level pairs or triplets. We fine-tune S-BERT model with a triplet-based contrastive learning approach. As illustrated in Fig. 1a–d, we create belief triplets from user voting activities, treating them as positive belief pairs, which are contrasted against negative examples. Specifically, we first create a vote co-occurrence dictionary, where each voted belief serves as a key, and the values consist of other beliefs that were also voted on by users who voted the key beliefs, allowing for duplication. From this dictionary, belief triplets are sampled (Methods). Notably, the more frequently two beliefs are shared by users, the more likely they are to be sampled as positive examples. Conversely, negative pairs are derived from beliefs that represent the opposing stance of the anchor belief or from beliefs that are frequently co-voted with the opposing belief.

These triplets are then utilized to fine-tune the LLMs using a triplet loss function as depicted in Fig. 1e,f. The resulting model offers a 768-dimensional latent belief space, where an individual belief is mapped into a vector within the space, and the distance between two vectors capture their semantic and contextual similarity. We show that the distance between beliefs in the belief space is proportional to the likelihood that an individual has one belief given their other belief.

We use two different approaches, a triplet evaluator and a semantic similarity evaluation task, to evaluate the quality of the belief embeddings generated by the LLMs. We initially assessed the performance of various LLMs by employing a triplet evaluator for classifying belief pairs as either positive or negative relations. Table 1 compares triplet evaluation results from different models. The fine-tuned S-BERT model showed the highest performance with an average accuracy of about 0.95 for the train dataset and about 0.67 for the test sets (Table 1).

Our model also shows good performance in capturing the general semantic meaning of various texts beyond the range of our training dataset, which is directly related to how accurate a vector representation of a new, unseen belief would be. Even after proceeding with the fine-tuning process, the S-BERT still retained a relatively high performance score on the GLUE-STSB[35] compared with other models. The Spearman rank correlation coefficient score of the S-BERT model is $r = 0.718 \pm 0.005$, while the fine-tuned BERT model shows a relatively low correlation score ($r = 0.476 \pm 0.045$) (Table 1).

Belief landscape revealed by belief embeddings

PCA results of the belief space. To examine the structure of the belief space, we perform principal component analysis (PCA) on the entire belief vectors generated from the fine-tuned S-BERT model. For analysis of the overall distribution of beliefs on various topics, we compiled 12 example sets of beliefs chosen from various fields that exhibit distinct patterns in the PC space, each consisting of a unique set of keywords relevant to their belief statements. For example, 5,000 distinct beliefs relate to the topic of 'God' and 1,470 beliefs relate to the topic of 'Gay marriage'.

Figure 2 presents the density of beliefs along the first and second principal component axes (PC1 and PC2) across belief subgroups, each related to distinct topics. The entire belief set (Fig. 2a) exhibits a smooth, uni-modal, bell-shaped distribution along both the first and second principal component axes (PC1 and PC2). However, the density plots for beliefs related to specific topics, such as 'God' and 'Abortion,' reveal markedly different, polarized patterns (Fig. 2b,c), suggesting that beliefs regarding these topics are grouped into two clusters with contrasting opinions. These bimodal patterns of beliefs are also observed in belief spaces using other types of dimensionality reduction methods (see Supplementary Fig. 9 for results using Uniform Manifold Approximation and Projection (UMAP)[36]).

Belief embeddings also reveal which beliefs are more closely associated with each other. For instance, beliefs favouring the existence of God (god) or opposing abortion are predominantly found on the negative side of the PC1 axis. The positive side of this axis is associated with disbelief in God and support for abortion rights. Additionally, beliefs related to topics, such as 'Gay and gay marriage', 'Evolution and Darwin' and 'Drug, Marijuana and Cannabis' also exhibit two dense clusters in the PC1 and PC2 space (Fig. 2d), which aligns with the broader trend of political polarization observed across diverse social issues[37–39].

Examining the other PC axes reveals another dimension of belief separation. For instance, beliefs about 'PlayStation' exhibit bimodal distributions along the PC2 axis, while beliefs related to 'Alcohol' display a weakly bimodal distribution along the PC3 axis (Fig. 2d). Similarly, beliefs about 'Harry Potter' and 'Twilight' cluster into two distinct groups on the PC3–PC4 plane, whereas they do not display a noticeable pattern on the PC1–PC2 plane. This indicates that the contextual relationships among beliefs concerning these topics are encoded in the PC3 and PC4 axes. However, not all topics show such polarized distributions; for example, beliefs related to 'Society', 'Education' and 'USA' tend to spread around in the PC space. This is probably because these topics encompass a wide array of subtopics, leading to greater variation in belief representation.

Overall, our results demonstrate that the distributions of beliefs related to various topics show unique patterns in the belief space, often forming polarized clusters along specific axes. Furthermore, the

Table 2 | Performance of various LLMs in a downstream task on predicting users' beliefs for unseen debates

Model type	Accuracy	Macro F1 score
S-BERT (fine-tuned)	0.590 (0.006)	0.590 (0.005)
S-BERT (before fine-tuning)	0.565 (0.002)	0.527 (0.002)
BERT (fine-tuned)	0.579 (0.002)	0.578 (0.002)
BERT (before fine-tuning)	0.541 (0.001)	0.496 (0.001)
Baseline 1 (random choice)	0.499 (0.002)	0.499 (0.002)
Baseline 2 (majority selection)	0.532 (0.001)	0.347 (0.001)
Llama2-13b-chat	0.537 (0.002)	0.371 (0.002)

Numbers in parentheses denote standard deviations obtained from five-fold validation results.

arrangement of belief positions of various polarizing issues in the PC space is generally aligned with the partisan polarization observed in public surveys. For instance, according to Gallup's beliefs poll in 2019[37], American liberals were more likely to consider 'Abortion' (73%) and 'Gay/lesbian relations' (81%) morally acceptable. In contrast, only 23% and 45% of conservatives believed these issues to be morally acceptable, respectively, demonstrating the interconnected nature of these beliefs.

Embedding individuals in belief space reveals group polarization. The presence of topic-specific bimodal belief distributions leads to a question: do individuals with distinct ideologies also exhibit meaningful clusters within the belief space? To investigate this, we represent each user by their average belief vector, defined as $\mathbf{u} = \sum_{i=1}^{N_u} \mathbf{b}_i^u / N_u$, where $\mathbf{b}_i^u$ denotes the ith belief vector of user u, and N_u is the total number of beliefs associated with user u. This allows us to measure how closely users are positioned in the belief space.

We then visually map users in the belief space according to their self-reported survey responses to assess whether the resulting user representations properly locate them. In DDO, users can self-report their positions on major social issues via pre-survey participation, independently from their debate participation. This includes specifying their supporting political parties, religious beliefs and positions on 48 key social issues, referred to as big issues. The big issues encompass a range of controversial social issues such as 'Abortion', 'Drug legalization', 'Gun control' and others.

Figure 3a–d illustrates the positions of users in the belief space along the first two PC axes. These positions are obtained by averaging their belief embeddings from S-BERT models before and after fine-tuning. The colour coding in these figures reflects the users' self-reported political ideologies (that is, Democrat versus Republican) and religious ideologies (that is, Christian versus Atheist). Remarkably, users represented by their average beliefs derived solely from voting records form two distinct clusters corresponding to their political and religious ideologies.

Figure 3b,d, which depict results from the fine-tuned S-BERT model, show a notable separation of user groups along the PC1 axis, suggesting that this axis primarily captures the alignment of users' beliefs with political and religious ideologies. By contrast, the base S-BERT model without fine-tuning does not exhibit such clear ideological group separations (Fig. 3a,c and Supplementary Fig. 10). This demonstrates that the fine-tuned S-BERT model more effectively captures the contextual relationships between beliefs, positioning related beliefs closer within the space.

The fine-tuned model also effectively reveals the alignment of partisan identity with beliefs on other disparate social issues. As shown in Fig. 3e, distinct user groups on various issues, such as 'Gay marriage', 'Abortion', 'Euthanasia' and 'Global warming exists', exhibit separation along the same PC1 axis, which represents partisan polarization. However, user groups are less clearly separated on other issues, such as 'Smoking ban' and 'Affirmative action'. This may be because these issues do not align neatly with prominent political or social dichotomies, such as the liberal–conservative spectrum, and because complexities beyond such dichotomies may not be fully captured by the PC1 axis. For example, individuals from both liberal or conservative backgrounds might either support or oppose a smoking ban. Similarly, perspectives on affirmative action may be influenced more by ethnicity than by political affiliation.

We further examine how the Euclidean distance between the PRO and CON user group centroids correlates with self-reported partisan polarization across the 48 big issues (Supplementary Section 4E and Supplementary Fig. 11). The analysis reveals a significant positive correlation ($r = 0.627$, $P < 0.001$), indicating that distances in the belief space accurately reflect the intensity of ideological polarization across these issues.

We note that our results primarily reflect user behaviours within the US-centric DDO dataset, and the observed clustering and polarization patterns may differ in other social and cultural contexts. Nevertheless, our findings demonstrate that the belief embedding framework effectively captures meaningful contextual relationships across diverse societal beliefs.

Belief embedding predicts user beliefs on unseen debates

Our results show that like-minded individuals with similar beliefs on specific social issues tend to cluster together in the belief space. This observation raises two further questions: Can we utilize the user embeddings to predict an individual's belief on unseen debates? Can we uncover any underlying mechanisms of human decision-making by analysing large-scale data on how users select their beliefs? According to the literature on dissonance theory of human attitudes, people tend to experience cognitive dissonance when they are exposed to information that is not in alignment with their existing beliefs[14,40]. Moreover, the feeling of personal discomfort created by the conflict between the new information and one's own beliefs can possibly lead to selective exposure to belief-confirming information[41]. Similarly, in our study, we consider a user's prior beliefs about various debate issues to constitute their existing belief system. Choosing a new belief towards a new debate is akin to adding a new belief to a user's existing belief system.

To quantitatively model the belief selection process, we design a binary belief classification task to predict a user's voting position (PRO or CON) in new debates. For this, we split the entire set of debates into an 8:2 ratio and evaluate the model's performance using fivefold cross-validation, considering users who appear at least once in both the training and test sets (Supplementary Table 4). We leverage user embeddings, learned from the training set, to predict a user's positions on previously unseen debates from the test set. We compare our results with multiple baselines and existing models.

Our model employs a straightforward approach; it predicts a user's choice on the basis of the Euclidean distance between the user's position and two opposing belief vectors from a new debate. For each debate in the test set, we construct two opposing belief statements from the debate title and generate their corresponding belief vectors, $\mathbf{b}_{\mathrm{PRO}}$ and $\mathbf{b}_{\mathrm{CON}}$, using the fine-tuned S-BERT model (Fig. 4a). Given a user embedding $\mathbf{u}$, representing the average of their prior beliefs, we compute the distances $d(\mathbf{u}, \mathbf{b}_{\mathrm{PRO}})$ and $d(\mathbf{u}, \mathbf{b}_{\mathrm{CON}})$, and predict the user's choice as the belief vector minimizing the distance: $\mathbf{b}' = \arg\min_{\mathbf{b} \in \{\mathbf{b}_{\mathrm{PRO}}, \mathbf{b}_{\mathrm{CON}}\}} d(\mathbf{u}, \mathbf{b})$.

Comparative evaluation with other LLMs reveals that the fine-tuned S-BERT model exhibits the highest performance, with an F1 score of 0.59 ($\sigma = 0.01$) and an accuracy of 0.59 ($\sigma = 0.01$). We use the macro F1 score to ensure balanced performance evaluation across all classes. This performance is notably superior compared with other models, including the base S-BERT model, the base and fine-tuned BERT models, and other baseline models, as presented in Table 2.

We also benchmark our model against two baseline models: the random choice model (baseline 1) and the majority selection model (baseline 2). The random choice model randomly predicts a user's belief between two given belief options. By its random nature, it is

https://doi.org/10.1038/s41562-025-02228-z

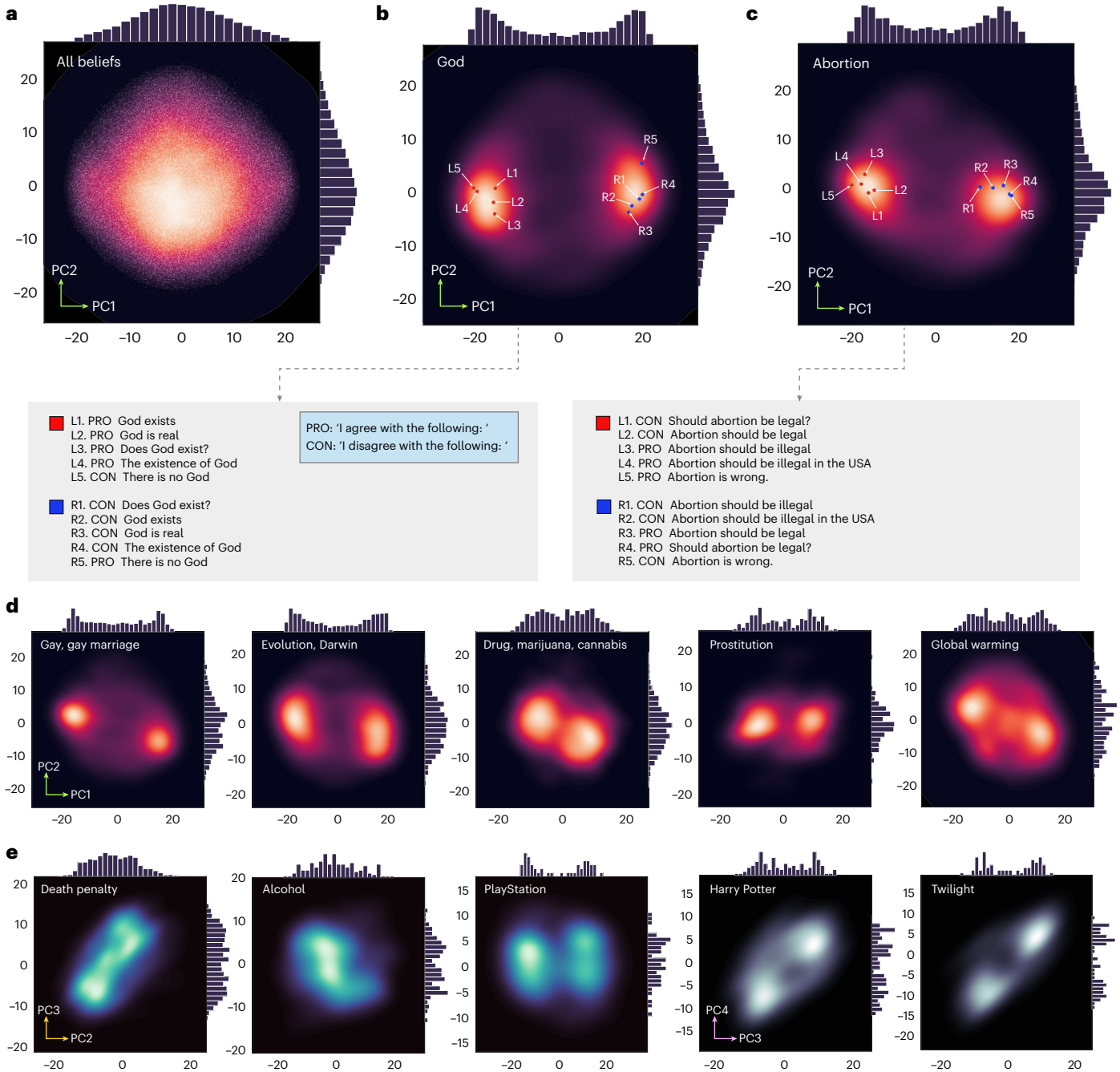

Fig. 2 | Structure of belief space via PCA. a, The entire distribution of belief embeddings represented in the first two principal components (PC1–PC2) space. The background heatmap reflects the density of beliefs, and the overlaid white dots indicate individual beliefs. **b**,**c**, The distribution of beliefs related to the topics 'God' (**b**) and 'Abortion' (**c**). Both belief distributions exhibit two highly clustered regions in the PC space, signifying the presence of two distinct groups of beliefs regarding these topics. Five example beliefs from each cluster, presented in the grey boxes, are highlighted as red and blue points. **d**, The distributions of beliefs corresponding to five different topics, displaying two bimodal distributions in the first two principal components (PC1–PC2) space, similar to **a** and **b**, are illustrated. **e**, Additional belief distributions corresponding to five topics, revealing unique structures in the higher-order principal components, are displayed.

expected to achieve an F1 score and accuracy of 0.5. The majority selection model accounts for the asymmetric ratio of PRO and CON beliefs in the training set by predicting that all users will consistently choose the more prevalent side. The majority model registers higher accuracy (0.53) but a lower F1 score (0.35). The fine-tuned S-BERT model outperforms both of these baselines. Additionally, we evaluate Llama2 (Llama2-13b-chat)[42] in a few-shot setup (Methods), achieving an accuracy of 0.54 and an F1 score of 0.37, slightly outperforming the majority baseline (Table 2).

Although the fine-tuned S-BERT model shows superior performance in belief prediction compared with other models, its overall F1 score is not particularly high (0.59). To understand the intricacies affecting the performance of the belief prediction, we explore various factors and identified four critical ones: the length of individuals' voting history, debate category, individual's effective radius and the relative distance between a user and two beliefs being considered.

First, our findings indicate that the prediction accuracy largely depends on the extent of a user's voting history in the training set.

https://doi.org/10.1038/s41562-025-02228-z

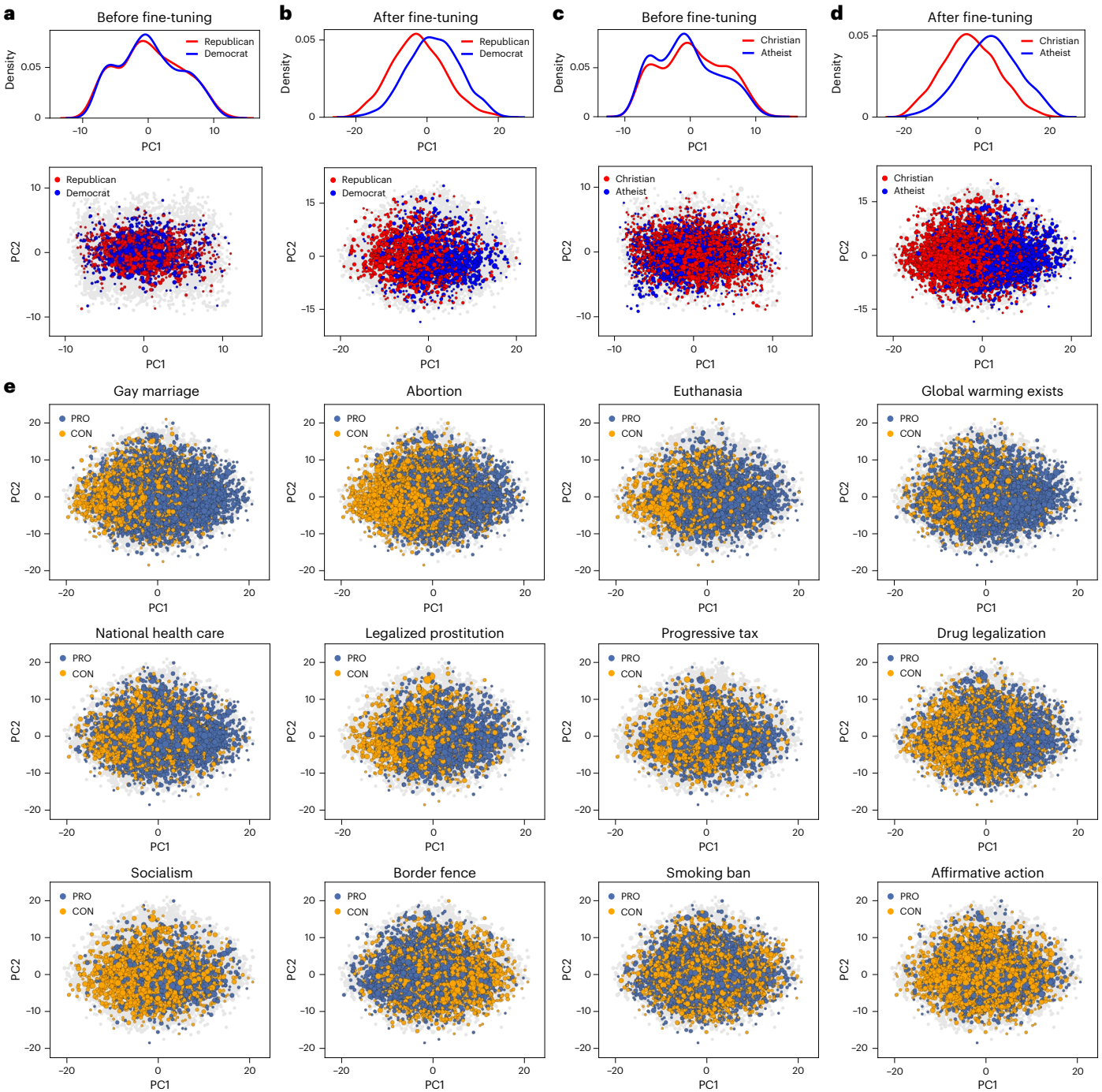

Fig. 3 | User embeddings in belief space. a–d, The positions of every user in the belief space along the first two principal component (PC) axes, coloured based on their self-reported supporting political parties (Republican versus Democrat, **a** and **b**) and religious ideologies (Christian versus Atheist, **c** and **d**), illustrating user embeddings obtained by averaging their belief embeddings from the base S-BERT model without fine-tuning (**a** and **c**) or from the fine-tuned S-BERT model (**b** and **d**). Compared with the non-fine-tuned model, the fine-tuned model exhibits a clearer separation of users from different groups in the belief space. **e**, Grouping of users by their self-reported stances on various controversial social issues.

Figure 4b and Supplementary Fig. 13 show that, as we progressively include users with longer voting history, the average F1 score of users almost monotonically increased. This result shows that the users' beliefs are more accurately predicted when we know more about their prior beliefs. However, a fundamental obstacle to accurate prediction is that the degree of user participation activities in DDO follows a highly skewed distribution (Fig. 4c), which is commonly found in many online human activities[43].

Second, the diverse nature of debate topics also poses a challenge. While debates related to politics and religion are common, many debates in the DDO dataset focus on issues closely tied to pop culture and recreational topics, such as 'Batman could beat Spiderman in a fight' or 'Soccer as the best sport'. Beliefs on such topics are often highly distinct from those on other issues, complicating predictions unless the user has previously engaged in similar topics. We utilize the topic categories provided in the DDO dataset and measured prediction performance across different debate categories. As highlighted in Fig. 4d, the prediction performance varies considerably over debate topics. For instance, the user beliefs related to 'religion' and 'philosophy' are

https://doi.org/10.1038/s41562-025-02228-z

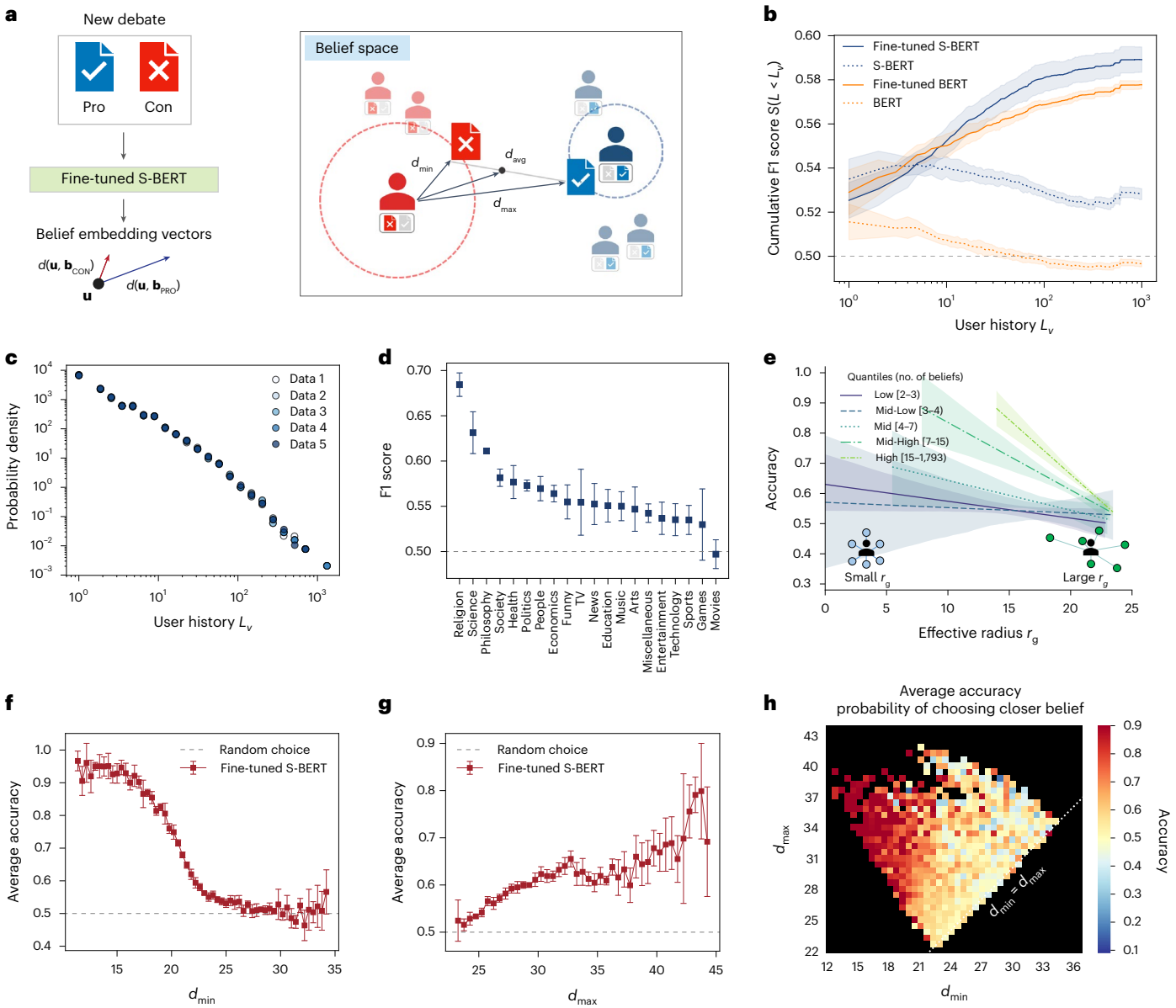

Fig. 4 | Predicting users' beliefs about new debates. a, For the two competing debate positions, PRO and CON, in an unseen debate, we define their belief vectors as $\mathbf{b}_{PRO}$ and $\mathbf{b}_{CON}$, respectively. Given a user's prior belief $\mathbf{u}$, we compute the distances $d(\mathbf{u}, \mathbf{b}_{PRO})$ and $d(\mathbf{u}, \mathbf{b}_{CON})$. A model predicts the user's choice as the belief vector that minimizes the distance, given by $\hat{\mathbf{b}} = \arg\min_{\mathbf{b}\in\{\mathbf{b}_{PRO},\mathbf{b}_{CON}\}} d(\mathbf{u}, \mathbf{b})$. The minimum and maximum distances are denoted as d_{min} and d_{max}, respectively, with their average labelled as d_{avg}. **b**, The relationship between the F1 score $S(L < L_v)$ and the length of user history L_v, where $S(L < L_v)$ represents the F1 score for users with voting records shorter than L_v. Data are presented as mean ± s.d. across five-fold cross-validation. **c**, The distribution of user history L_v (number of votes) in a fivefold dataset. **d**, The variation in belief prediction accuracy, quantified using the F1 score, across diverse debate categories. The graph depicts top 20 most frequent categories, each appearing over 100 times in the prediction task. Data are presented as mean ± s.d. **e**, The accuracy trend as a function of users' effective radius. The lines represent user groups divided into five quantiles on the basis of the number of prior beliefs, with each group containing a similar number of users. The shaded areas indicate the 95% confidence interval around the fitted mean regression line. **f**,**g**, The average accuracy trends across d_{min} (**f**) and d_{max} (**g**) in the belief prediction task. Error bars represent mean ± s.d. **h**, A heatmap illustrating the average prediction accuracy across various ranges of d_{min} and d_{max}.

more predictable than those under 'sports', 'funny' and 'games'. This discrepancy remains consistent even after downsampling the dominant categories, resulting in a training dataset with a relatively more balanced distribution of debates across categories (Supplementary Section 6 and Supplementary Fig. 20).

Third, users can exhibit varying distributions of beliefs within the belief space. Despite having the same number of prior beliefs (voting history), some users display a broader distribution of beliefs, while others show more concentrated beliefs in a smaller region. To investigate how belief selection patterns differ between these groups, we quantify the dispersion of a user's prior beliefs using a metric called the effective radius (r_g). This measure, analogous to the radius of gyration, is defined for a user u as

$$r_g^u = \sqrt{\sum_{i=1}^{N_u} \| \mathbf{b}_i^u - \mathbf{u} \|^2 / N_u}, \quad (1)$$

where $\mathbf{b}_i^u$ denotes the ith prior belief vector of user u and $\mathbf{u}$ represents the centroid of the user's prior beliefs. Here, N_u is the total number of prior beliefs of user u.

Figure 4e illustrates the relationship between effective radius and average belief prediction accuracy, where users are grouped into five

https://doi.org/10.1038/s41562-025-02228-z

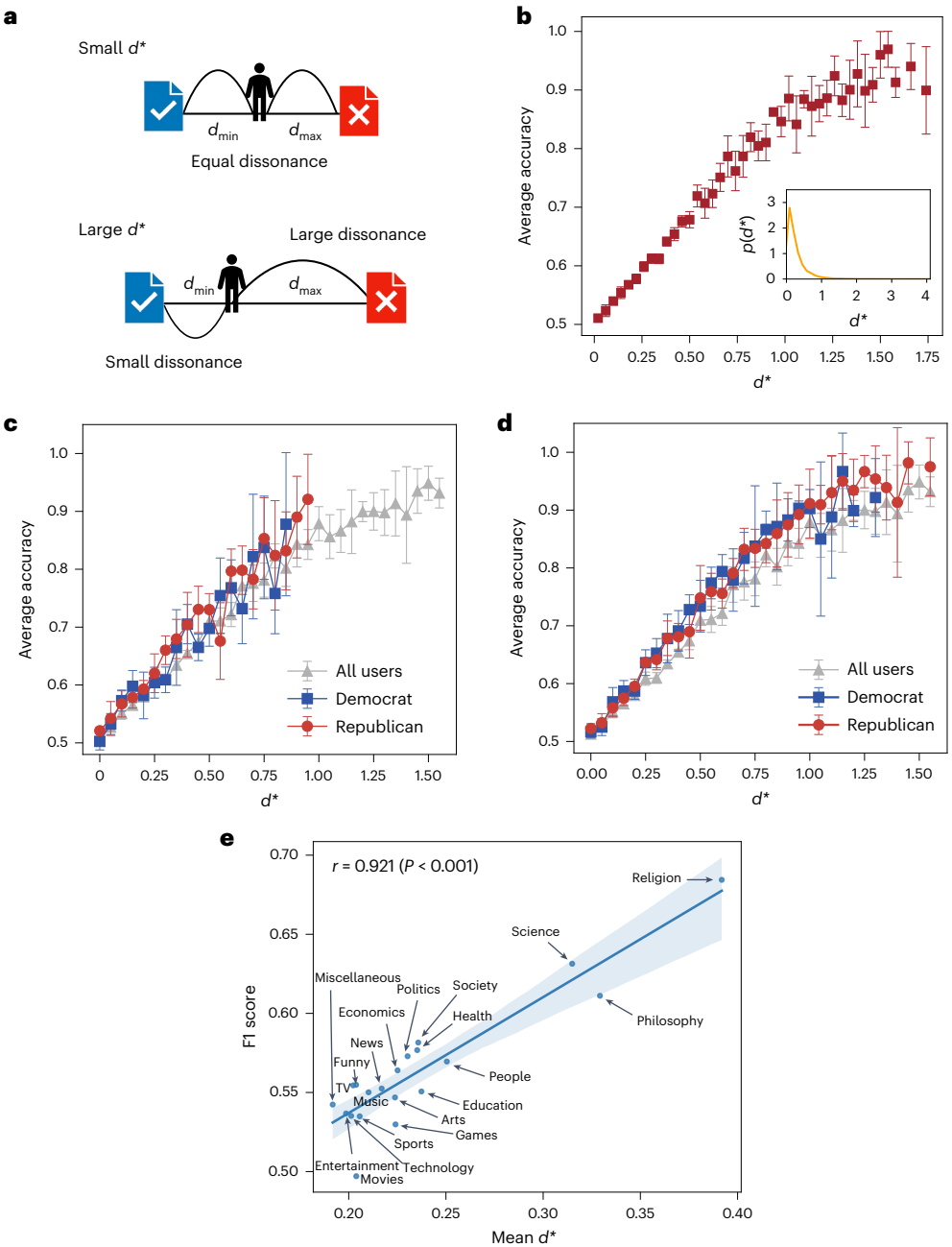

Fig. 5 | Effect of relative dissonance on belief selection. a, An illustration of two scenarios in which a user selects a belief for a debate, contrasting cases of small and large relative dissonance. When the two beliefs under consideration are equally distant from the user, selecting either belief may result in an equal level of dissonance. By contrast, when one belief is much farther than the other, the potential dissonance a user may experience varies depending on their selection. **b**, The likelihood of a user choosing a closer belief linearly increases with relative dissonance d^*. The inset shows the distribution of d^*. **c**,**d**, The linear relationships between average accuracy in the belief prediction task and d^*, shown separately for two user groups, namely, Democrats versus Republicans (**c**) and Christians versus Atheists (**d**). The results show a similar linear relationship regardless of users' political or religious ideologies. Error bars in **b**–**d** represent the s.d. across the results of the fivefold validation tasks. **e**, The average d^* for different debate topics within the prediction task and the corresponding average prediction score across topic areas. Topics with higher mean d^* tend to have more accurate predictions. The solid regression line represents this trend, with the shaded area indicating the 95% confidence interval.

quantiles on the basis of their number of prior beliefs. Across all user groups, belief prediction accuracy decreases as r_g increases, indicating that individuals with more concentrated beliefs (smaller r_g) are more likely to select a belief closer to their prior ones between two opposing beliefs in an unseen debate. By contrast, users with more dispersed beliefs are more likely to select the belief farther from their prior beliefs. It is important to note that, since our model always assumes that users will choose a belief closer to their prior belief, prediction accuracy can be directly interpreted as the probability of a user choosing the closer belief. Therefore, for the remaining analyses, we use the accuracy score instead of the F1 score for clearer interpretation, a decision further supported by the fact that the two scores are highly correlated.

Fourth, the proximity of a user to the beliefs under consideration in the belief space substantially impacts prediction accuracy (Fig. 4f–h). When two opposing beliefs are introduced in a new debate during the prediction task, we measure their distances from a user: d_{min} for the

closer belief and d_{max} for the farther belief. Our analysis shows that the prediction accuracy is inversely correlated with d_{min} and positively correlated with d_{max} (Fig. 4f,g). The heatmap in Fig. 4h illustrates how the average accuracy varies across the two-dimensional space defined by d_{min} and d_{max}. This suggests that the probability of choosing the closer belief decreases as d_{min} increases, and increases as d_{max} increases. Consequently, the predictions are more accurate when the closer belief is much closer to the user and the farther belief is much farther away.

We further assess the impact of the average distance d_{avg} between the user and two opposing beliefs on the prediction accuracy (Supplementary Fig. 14). As d_{avg} increases, the accuracy converges to 0.5, equivalent to random guessing. This suggests that, when both beliefs are sufficiently distant from the user's position (when $d_{avg} \approx 33$), predicting the user's choice becomes extremely difficult. A large d_{avg} indicates that the debate introduces viewpoints that are distant from or weakly associated with the user's prior beliefs, consequently reducing the predictive power of past voting behaviour. We conducted comprehensive robustness checks under various data splitting and testing scenarios, confirming the consistency of our findings across different conditions (Supplementary Section 6).

In addition to analysing these factors, we examined whether beliefs varied in predictability across user groups based on political party, religion or sex. However, we found no significant differences in predictive accuracy across these groups (Supplementary Fig. 15).

Role of relative dissonance in belief prediction

Our findings from the belief prediction task reveal that the distances between a user and two opposing beliefs under consideration (d_{min} and d_{max}) substantially impact prediction accuracy. The underlying patterns of belief selection provide important insights into human decision-making mechanisms.

On the basis of these empirical observations, we propose a new metric, termed 'relative dissonance', denoted as d^*, to better quantify this decision-making process

$$d^* = \frac{d_{max} - d_{min}}{d_{min}}. \quad (2)$$

d^* represents the absolute difference between a user's distances to two opposing beliefs, normalized by the shorter distance, d_{min}.

From the perspective of cognitive dissonance, a belief closer to a user's existing beliefs (that is, smaller d_{min}) is expected to result in less dissonance, while a belief that is farther away would induce greater dissonance (Fig. 5a). Thus, d^* serves as a relative measure of dissonance reduction when a user opts for the belief closer to their prior beliefs rather than a more distant belief.

We note that we use the term 'dissonance' in a broad sense to represent the distance between a user and a belief. In our framework, a user vector is computed by averaging their belief vectors, capturing the average position of the beliefs held by the user. The distance between a user vector and a new belief vector reflects, on average, how much the new belief deviates from (or is dissonant with) the user's prior beliefs. In this context, dissonance is treated as a continuous measure, analogous to distance. Smaller dissonance values indicate alignment or favourability towards the belief, whereas larger dissonance values reflect greater deviation from the user's prior beliefs.

Figure 5a,b illustrates a compelling relationship between d^* and the average prediction accuracy of users' beliefs on new debates: the average accuracy shows a linear increase with the rise of d^*. As we noted previously, the prediction accuracy can be interpreted as the probability that a user selects a belief closer to their prior beliefs. Thus, the observed linear increase suggests that this probability depends on relative dissonance. In other words, when the potential dissonance from one belief outweighs that from another, users are more likely to choose the belief that is closer to them.

When d^* is near 0, the probability of a user choosing a closer belief is around 50%, suggesting that their decisions are largely independent of their prior beliefs (Fig. 5b). Conversely, for debates when d^* is larger (for example, $d^* \approx 1.5$), users exhibit a strong preference for beliefs closer to their position, with probability close to 1. In this scenario, users are strongly inclined to select beliefs closely aligned with their prior ones, indicating a substantial reduction in potential dissonance when avoiding an alternative belief.

We further investigate whether user groups with distinct political or religious ideologies—specifically, comparisons between Democrats and Republicans, as well as Christians and Atheists—exhibit different decision-making patterns with respect to relative dissonance d^*. As shown in Fig. 5c,d, we find no significant difference in how relative dissonance influences belief selection between two groups ($P > 0.05$ in two-sample t tests across all d^* ranges). These results suggest that the impact of relative dissonance on belief selection is remarkably consistent across different political and religious groups.

The concept of relative dissonance also helps to explain why users' beliefs on certain debate topics are more predictable than others. We found a strong correlation ($r = 0.921$) between the average d^* of a debate category and its prediction F1 score (Fig. 5e). This high correlation partly explains why users' belief choices on certain topics (for example, 'religion' or 'philosophy') are more predictable than in others (for example, 'funny' or 'entertainment'). For example, the difference in distance from a user to two opposing beliefs tends to be much larger in the belief space for debates on 'religious' topics than for those on 'funny' or 'entertainment.'

To assess whether the correlation between a category's average d^* and prediction accuracy is merely a byproduct of category size, measured by the number of training examples per category, we analyse the relationships among mean d^*, average F1 score, and category size (Supplementary Section 6E). Our robustness checks indicate that while d^* and category size are partially correlated, d^* consistently exhibits a stronger and more robust correlation with prediction accuracy than category size. This trend becomes even more pronounced when frequently occurring debate categories are downsampled.

Discussion

Here, we demonstrate that neural embedding approaches based on LLMs offer a powerful and scalable solution for understanding the complex and nuanced relationships among human beliefs. While previous approaches provide insightful theoretical bases for modelling belief systems that incorporate belief relationships, there has been a lack of robust frameworks to comprehensively represent the space of beliefs encompassing a wide range of topics[14,15,24]. Existing methods, which often rely on surveys and small, topic-specific datasets, lack scalability and face challenges in capturing the full spectrum of beliefs individuals hold.

In this perspective, LLMs integrated with user activity data can open a new avenue for modelling human beliefs. Pre-trained language models, which already possess a strong understanding of complex language patterns and contextual information, can be fine-tuned using extensive belief records to create a comprehensive 'embedding space of human beliefs'. This embedding space maps a wide range of topics and enables inductive reasoning about new beliefs. Furthermore, this approach efficiently represents an individual's belief system and supports various downstream tasks such as quantifying polarization or predicting beliefs.

The key findings from our study offer several insights into the characterization of human beliefs. First, our study introduces a representative learning framework for constructing a belief embedding space in a continuous high-dimensional vector space using online user activities and LLM. This space effectively reveals the interconnected structure of various human beliefs and the polarization of beliefs related to representative social issues. The continuous belief space

created using the fine-tuned LLM facilitates inductive reasoning, enabling the addition of new beliefs.

Second, the vector representation of individuals allows us to identify how people with different opinions are clustered and polarized. The fine-tuned S-BERT model reveals a clearer separation among individuals with similar political or religious ideologies, whereas the base S-BERT model without fine-tuning does not exhibit such patterns. Our results demonstrate the usefulness of the belief space in measuring the polarization of certain social concepts. The distance between groups of individuals with opposing beliefs on a given issue within the belief space is highly correlated with the degree of political polarization associated with that issue.

Third, the downstream task for belief prediction shows that the proposed belief space is useful for predicting individuals' beliefs on new debates on the basis of their pre-existing beliefs. We uncover four critical factors that influence the prediction outcome of an individual's choice of a new belief: the length of individuals' voting records, debate categories, effective radius of individuals, and the distances between the individual and the two beliefs under consideration in the belief space.

Most importantly, our empirical observations highlight that the relative distance between an individual and two opposing beliefs in a new debate is a reliable predictor of their decision. This insight lead us to develop a novel metric called 'relative dissonance' d^*, which quantifies the relative inconsistency a person may experience when adopting a belief into their pre-existing belief system compared with its opposite belief (Fig. 5). Our analysis reveals that, as the relative dissonance (d^*) increases, the likelihood of a person choosing a belief closer to their current position in the belief space increases. In other words, the greater the difference in dissonance a person experiences between two beliefs, the more likely they are to choose the belief that causes less dissonance. This finding aligns with conventional cognitive dissonance theory and offers a quantitative measure of cognitive dissonance by linking it to distances within the belief space.

While our model captures many aspects of human belief dynamics, our study does have limitations that will guide future research. First, the reliance on a single online debate platform for data collection may limit the generalizability of our findings. Incorporating broader datasets from diverse platforms will help understanding the universal properties of belief systems and their cultural and social variations. Additionally, the dataset used in this study is primarily based on US data, which may not fully represent global perspectives and cultural diversity in human beliefs. Future research should include data from various societies to achieve a broader relevance of the findings across different cultural contexts.

Second, the DDO dataset used in this study, where users' preferences are easily inferred from explicit voting records, represents a specific data type. Developing methods for extracting human beliefs from more general texts on diverse platforms, such as social media postings, news interviews, and movie scripts, would provide a deeper understanding of human beliefs and increase the applicability of our framework.

Third, our study does not investigate the temporal and dynamic properties of the belief space. Although our study indirectly assumes the stability of the belief space, in reality, a society's beliefs on social issues can continuously change. Investigating how the shape of the entire belief space, which reflects the interconnections of collective societal beliefs, transforms over time would be an interesting avenue for future research.

Fourth, there is a concern regarding the inherent biases present in the pre-trained LLMs used in our study[44]. For example, LLMs trained predominantly on English-language internet data may inadvertently reflect Western-centric viewpoints, underrepresenting or misrepresenting beliefs prevalent in non-Western cultures. These models might exhibit biases related to sex, race and socioeconomic status, which could skew the analysis of the belief relationships. Ongoing efforts to improve fairness and reduce biases in LLMs are crucial for future research to ensure more equitable and accurate representations of human beliefs.

Looking ahead, while our primary goal in this study is to create a comprehensive map of beliefs and uncover the mechanisms behind human belief selection, our contrastive learning approach also shares certain core principles with recommendation systems[45–47]. We anticipate that our contrastive learning methods—extracting both positive and negative relationships from user activities as well as utilizing the semantic understanding of LLMs—could be effectively applied in recommendation algorithms. Moreover, integrating the cognitive patterns and belief dynamics revealed in this study may enable recommendation systems to better reflect how human beliefs evolve and interact, ultimately leading to more personalized and context-aware suggestions.

In essence, our research establishes a foundational framework for an advanced, data-driven analysis of human beliefs using LLM. We anticipate that this work on the complex landscape of human beliefs would provide both theoretical insights and practical applications in understanding and modelling human behaviour in the fields of cognitive science, social psychology, political science and beyond.

Methods

DOO dataset and extraction of belief statements

The DDO dataset used in this study contains a corpus of 78,376 debates (68,900 unique debate titles excluding duplicates) by 42,906 debaters from 15 October 2007 to 19 September 2018 (Supplementary Figs. 1 and 2). In DDO, each debate features two debaters, one supporting the proposition (PRO) and the other opposing it (CON). In each debate, other users can engage by voting on seven different items. Notably, the option 'Agree with after the debate' enables users to express their position on the debate topic as either PRO, CON, or TIE, reflecting their belief on the issue. To extract belief pairs that reveal clear positive and negative relations, we only considered the PRO and CON votes and excluded TIE votes. We also treated debaters and voters equally as voters, as our study utilizes users' positions on various debate topics as their beliefs.

Most debate titles in DDO represent beliefs on various topics (for example, 'Abortion should be legal', 'God exists' and 'All morals are relative'). Thus, users' votes on these titles as PRO or CON can be considered as revealing their beliefs on these topics. To generate a complete belief statement for a user, we appended a template phrase that explicitly describes the user's stance. For example, a PRO (or CON) vote on a debate title leads to the belief statement, 'I agree (disagree) with the following: [DEBATE TITLE]'. For instance, a PRO vote on 'Abortion is morally justified' results in the belief statement, 'I agree with the following: abortion is morally justified'. These belief statements are then fed into LLMs.

We performed data filtering on the DDO dataset to make it suitable for our analyses. While most debate topics in DDO can be considered in the form of beliefs that allow for support or opposition, there are also incomplete or unsuitable titles that cannot be regarded as beliefs. We filtered these unsuitable debate titles using GPT-4 (refs. 48,49), one of the most advanced and reliable artificial intelligence language models at the time of our study. We asked GPT-4 to determine whether a given statement (debate title) can be considered a human belief (Supplementary Section 2, Supplementary Table 1 and Supplementary Figs. 3 and 4).

Among 68,900 unique debate titles, GPT-4 classified 8,914 as unsuitable for consideration as belief statements. The unsuitable debate titles include titles that use 'versus' or 'vs.', such as 'Batman versus Spiderman' and 'atheism versus agnosticism,' titles denoting battle content such as 'Rap battle', 'music battle' and 'Video Rap battle', titles with single words without meaningful context or incomplete sentences, for instance, 'fox news', 'useless', 'Media are …', titles posing 'how' questions such 'How many donuts are too many donuts', 'How

can you be an atheist?' as well as titles expressing personal resolutions or suggestions such as 'I will not contradict myself' and 'I will lose this debate'. Removing 8,914 inadequate debate titles resulted in 59,986 unique debate titles (from 65,861 debates) that were voted on a total of 192,307 times by a total of 40,280 users.

To assess the consistency of classification results using GPT-4 with human annotations, we compared its classifications of 50 randomly sampled debate titles against those determined by three human annotators (three of the authors on this study). We equally sampled 25 titles from each category of 'True' and 'False', as classified by GPT-4, to ensure balanced representation. The annotators were requested to indicate whether or not the debate titles qualify as belief statements. The inter-annotator reliability, measured using Fleiss' Kappa—which quantifies agreement beyond chance—was 0.866, indicating a high level of agreement among the human annotators. GPT-4's classifications showed an 88% agreement rate with the majority vote of the human annotators. This high agreement rate suggests the strong consistency between GPT-4's classifications and the consensus among human annotators in identifying belief statements.

Training LLMs with belief triplets to build belief space

We employed a pre-trained S-BERT model (roberta-base-nli-stsb-mean-tokens)[32] on the basis of the RoBERTa model[33], to learn relationships between beliefs across multiple topics. Using belief triplets, we applied a contrastive learning technique to fine-tune the model. We explored various LLMs, from the original BERT[34] to other S-BERT models pre-trained with different sources. The RoBERTa-based model exhibited superior performance in diverse tasks and was thus selected for our study.

For the fine-tuning process, we created belief triplets using the voting records of users. A user's voting records on various debates create a sequence of beliefs. Using these belief sequences, we produced a set of belief triplets. Each of the belief triplets comprises three distinct beliefs: an anchor belief statement B_a, a positive example belief B_p and a negative example belief B_n. We went through all belief statements as anchor beliefs and found corresponding positive and negative examples. The positive example beliefs for a given anchor were sampled from the beliefs that were voted on together with the anchor belief, weighted by their frequency (the more often two beliefs are voted on by the same users, the more likely they are to be sampled as positive examples). Conversely, the negative example beliefs of an anchor belief were selected from either the directly opposing belief statement (expressing an opposite opinion towards the anchor belief) or from the beliefs that were co-voted with the opposite belief statement.

For example, assume that many users frequently voted as PRO to the debates titled 'Abortion is morally justified' and 'Same-sex marriage should be legal'. Then, for the anchor belief, 'I agree with the following: abortion is morally justified', a possible positive example could be 'I agree with the following: same-sex marriage should be legal', and a negative example could be 'I disagree with the following: abortion is morally justified'. In this way, we sampled at most five positive examples and five negative examples for a given anchor belief statement, and generated all possible combinations of belief triplets on the basis of these examples. A maximum of 25 triplets can be created for one anchor belief.

We note that the same pair of beliefs may appear both as a positive and negative example in different proportions. For example, a belief pair could be co-voted together (positive) by a majority of users yet be opposed (negative) by a minority. By including all such variations, our model learns a weighted, continuous measure of similarity, enabling us to move beyond a simplistic binary determination of 'similar' versus 'dissimilar'.

The belief triplets were fed into the pre-trained S-BERT model. We divided debates into training and test data in an 8:2 ratio, repeating this process five times for fivefold validation datasets. On average, 1,354,123 triplets were used for the fine-tuning process as training sets. The model was fine-tuned to minimize the triplet loss function

$$L = \max(\| \mathbf{s}_a - \mathbf{s}_p \| - \| \mathbf{s}_a - \mathbf{s}_n \| + \epsilon, 0), \quad (3)$$

where $\mathbf{s}_a$, $\mathbf{s}_p$ and $\mathbf{s}_n$ are the 768-dimensional output vectors of S-BERT corresponding to the sentence embedding of an anchor belief B_a, a positive belief B_p and a negative belief B_n, respectively. ϵ is the triplet margin term, which guarantees that the negative belief vector $\mathbf{s}_n$ must be farther away from the anchor $\mathbf{s}_a$ than the positive belief vector $\mathbf{s}_p$. We used the default parameter $\varepsilon = 5$.

During training, the weight parameters of the S-BERT model are updated in order to minimize the Euclidean distance between $\mathbf{s}_a$ and $\mathbf{s}_p$, while simultaneously maximizing the gap between $\mathbf{s}_a$ and $\mathbf{s}_n$. The fine-tuned model thus provides a comprehensive 768-dimensional latent representation of human beliefs, termed the belief space. When belief statements are inputted into the LLM, it outputs their vector representations that form this belief space, where the positions and distances between beliefs reveal interdependencies between them.

Belief prediction with a larger LLM in a few-shot setting

During the downstream task, which involves predicting user beliefs on unseen debates, we benchmarked our results against the performance of Llama2 (Llama2-13b-chat)[42], a recent LLM with much larger parameters, for few-shot tasks. We chose Llama2 as it exhibits strong zero/few-shot performance across a variety of tasks such as question answering and natural language reasoning. For our task, Llama2 was prompted with a user's existing beliefs from the training set and tasked with predicting the user's stance on new, unseen debates. After testing several prompts, we chose a prompt for Llama2 that includes a user's prior belief statements, followed by a query: 'Based on these statements, do you think you might agree or disagree with the following: {DEBATE TITLE}? Please choose from one of these options: agree or disagree. Do not explain your choice'. This approach required the model to make a binary decision, answering either 'agree' or 'disagree'. We were able to test only on approximately 85% of the dataset owing to the context-size limitations of Llama2.

Ethics

Our study does not involve any human subjects or experiments and is not subject to institutional review board approval. Consequently, there were no ethical regulations to comply with, no informed consent was required and no participant compensation was involved. Additionally, our study was not preregistered.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The original DDO dataset[20,31] is available via GitHub at https://esdurmus.github.io/ddo.html. For the replication of our study, a processed version of this dataset, including pre-processed user-level debate records and the fine-tuned models used in our analyses, is available via GitHub at https://github.com/ByunghweeLee-IU/Belief-Embedding.

Code availability

We developed custom code using Python 3.9.10 for data analysis. The replication code is available via GitHub at https://github.com/ByunghweeLee-IU/Belief-Embedding.

References

1. *Cambridge Learner's Dictionary* (Cambridge Univ. Press, 2008).
2. *Oxford Advanced Learner's Dictionary* (Oxford Univ. Press, 2000).

3. DellaPosta, D., Shi, Y. & Macy, M. Why do liberals drink lattes? *Am. J. Sociol.* **120**, 1473–1511 (2015).
4. Goldberg, A. & Stein, S. K. Beyond social contagion: associative diffusion and the emergence of cultural variation. *Am. Sociol. Rev.* **83**, 897–932 (2018).
5. González-Bailón, S. et al. Asymmetric ideological segregation in exposure to political news on Facebook. *Science* **381**, 392–398 (2023).
6. Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W. & Starnini, M. The echo chamber effect on social media. *Proc. Natl Acad. Sci. USA* **118**, e2023301118 (2021).
7. Castellano, C., Fortunato, S. & Loreto, V. Statistical physics of social dynamics. *Rev. Mod. Phys.* **81**, 591–646 (2009).
8. DeGroot, M. H. Reaching a consensus. *J. Am. Stat. Assoc.* **69**, 118–121 (1974).
9. Sen, P. & Chakrabarti, B. K. *Sociophysics: an Introduction* (Oxford, 2014).
10. Deffuant, G., Neau, D., Amblard, F. & Weisbuch, G. Mixing beliefs among interacting agents. *Adv. Complex Syst.* **3**, 01n04, 87–98 (2001).
11. Friedkin, N. E. & Johnsen, E. C. Social influence and opinions. *J. Math. Sociol.* **15**, 193–206 (1990).
12. Watts, D. J. A simple model of global cascades on random networks. *Proc. Natl Acad. Sci. USA* **99**, 5766–5771 (2002).
13. Macy, M., Deri, S., Ruch, A. & Tong, N. Opinion cascades and the unpredictability of partisan polarization. *Sci. Adv.* **5**, eaax0754 (2019).
14. Galesic, M., Olsson, H., Dalege, J., van der Does, T. & Stein, D. L. Integrating social and cognitive aspects of belief dynamics: towards a unifying framework. *J. R. Soc. Interface* **18**, 20200857 (2021).
15. Aiyappa, R., Flammini, A. & Ahn, Y.-Y. Emergence of simple and complex contagion dynamics from weighted belief networks. *Sci. Adv.* **10**, eadh4439 (2024).
16. Dalege, J. et al. Toward a formalized account of attitudes: the causal attitude network (CAN) model. *Psychol. Rev.* **123**, 2–22 (2016).
17. Dalege, J. & van der Does, T. Using a cognitive network model of moral and social beliefs to explain belief change. *Sci. Adv.* **8**, eabm0137 (2022).
18. Rodriguez, N., Bollen, J. & Ahn, Y.-Y. Collective dynamics of belief evolution under cognitive coherence and social conformity. *PLoS One* **11**, e0165910 (2016).
19. Schweighofer, S., Schweitzer, F. & Garcia, D. A weighted balance model of opinion hyperpolarization. *J. Artif. Soc. Soc. Simul.* **23**, 5 (2020).
20. Durmus, E. & Cardie, C. Exploring the role of prior beliefs for argument persuasion. In *Proc. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (eds Walker, M., Ji, H. & Stent, A.) 1035–1045 (Association for Computational Linguistics, 2018).
21. Longpre, L., Durmus, E. & Cardie, C. Persuasion of the undecided: language vs. the listener. In *Proc. 6th Workshop on Argument Mining* (eds Stein, B & Wachsmuth, H.) 167–176 (Association for Computational Linguistics, 2019).
22. Agarwal, V., Joglekar, S., Young, A. P. & Sastry, N. GraphNLI: a graph-based natural language inference model for polarity prediction in online debates. In *Proc. ACM Web Conference* 2729–2737 (Association for Computing Machinery, 2022).
23. Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X. & Cherry, C. SemEval-2016 task 6: detecting stance in tweets. In *Proc. 10th International Workshop on Semantic Evaluation (SemEval-2016)* (eds Bethard, S. et al.) 31–41 (Association for Computational Linguistics, 2016).
24. Introne, J. Measuring belief dynamics on Twitter. *Proc. Int. AAAI Conf. Web. Soc. Media* **17**, 387–398 (2023).
25. Darwish, K., Stefanov, P., Aupetit, M. & Nakov, P. Unsupervised user stance detection on Twitter. *Proc. Int. AAAI Conf. Web. Soc. Media* **14**, 141–152 (2020).
26. Rashed, A., Kutlu, M., Darwish, K., Elsayed, T. & Bayrak, C. Embeddings-based clustering for target-specific stances: the case of a polarized Turkey. *Proc. Int. AAAI Conf. Web. Soc. Media* **15**, 537–548 (2021).
27. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient estimation of word representations in vector space. Preprint at https://arxiv.org/abs/1301.3781 (2013).
28. Le, Q. & Mikolov, T. Distributed representations of sentences and documents. In *Proc. International Conference on Machine Learning* (eds Xing, E. P. & Jebara, T.) 1188–1196 (PMLR, 2014).
29. An, J., Kwak, H. & Ahn, Y.-Y. SemAxis: a lightweight framework to characterize domain-specific word semantics beyond sentiment. In *Proc. 56th Annual Meeting of the Association for Computational Linguistics* (eds Gurevych, I. & Miyao, Y.) 2450–2461 (Association for Computational Linguistics, 2018).
30. Schroff, F., Kalenichenko, D. & Philbin, J. FaceNet: a unified embedding for face recognition and clustering. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 815–823 (2015).
31. Durmus, E. & Cardie, C. A corpus for modeling user and language effects in argumentation on online debating. In *Proc. 57th Annual Meeting of the Association for Computational Linguistics* (eds Korhonen, A., Traum, D. & Màrquez, L.) 602–607 (Association for Computational Linguistics, 2019).
32. Reimers, N. & Gurevych, I. Sentence-BERT: sentence embeddings using Siamese BERT-networks. In *Proc. 2019 Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (eds Inui, K., Jiang, J., Ng, V. & Wan, X.) 3982–3992 (Association for Computational Linguistics, 2019).
33. Liu, Y. et al. RoBERTa: a robustly optimized BERT pretraining approach. Preprint at https://arxiv.org/abs/1907.11692 (2019).
34. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C. & Solorio, T. (eds) *Proc. Annual Meeting of the Association for Computational Linguistics* 4171–4186 (Association for Computational Linguistics, 2019).
35. Wang, A. et al. GLUE: a multi-task benchmark and analysis platform for natural language understanding. In *Proc. 2018 EMNLP Workshop on BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (eds Linzen, T., Chrupała, G. & Alishahi, A.) 353–355 (Association for Computational Linguistics, 2018).
36. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: uniform manifold approximation and projection. *J. Open Source Softw.* **3**, 861 (2018).
37. Brenan, M. Birth control still tops list of morally acceptable issues. *Gallup* https://news.gallup.com/poll/257858/birth-control-tops-list-morally-acceptable-issues.aspx. (2019).
38. Gentzkow, M., Shapiro, J. M. & Taddy, M. Measuring polarization in high-dimensional data: method and application to congressional speech. *Econometrica* **87**, 1307–1340 (2019).
39. Campbell, J. E. *Polarized: Making Sense of a Divided America* (Princeton Univ. Press, 2018).
40. Festinger, L. *A Theory of Cognitive Dissonance* (Stanford Univ. Press, 1957).
41. Frimer, J. A., Skitka, L. J. & Motyl, M. Liberals and conservatives are similarly motivated to avoid exposure to one another's opinions. *J. Exp. Soc. Psychol.* **72**, 1–12 (2017).
42. Touvron, H. et al. LLaMA 2: open foundation and fine-tuned chat models. Preprint at https://arxiv.org/abs/2307.09288 (2023).

43. Muchnik, L. et al. Origins of power-law degree distribution in the heterogeneity of human activity in social networks. *Sci. Rep.* **3**, 1783 (2013).
44. Gallegos, I. O. et al. Bias and fairness in large language models: a survey. *Comput. Linguist.* **50**, 1–83 (2024).
45. Hu, Y., Koren, Y. & Volinsky, C. Collaborative filtering for implicit feedback datasets. In *Proc. 8th IEEE International Conference on Data Mining* 263–272 (IEEE, 2008).
46. Koren, Y., Bell, R. & Volinsky, C. Matrix factorization techniques for recommender systems. *Computer* **42**, 30–37 (2009).
47. Kim, S. et al. Large language models meet collaborative filtering: an efficient all-round LLM-based recommender system. In *Proc. 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* 1395–1406 (2024).
48. GPT-4. *OpenAI* https://openai.com (2024).
49. Achiam, J. et al. GPT-4 technical report. Preprint at https://arxiv.org/abs/2303.08774 (2023).

Acknowledgements

B.L. and Y.A. are supported in part by the Air Force Office of Scientific Research under award no. FA9550-19-1-0391. Y.A. was supported in part by DARPA under contract HR001121C0168. B.L., R.A., J.A., H.K. and Y.A. are in part supported by the Air Force Office of Scientific Research under award no. FA9550-25-1-0087. H.K. is supported by the Luddy Faculty Fellow Research Grant Programme of the Luddy School of Informatics, Computing and Engineering at Indiana University Bloomington.

Author contributions

B.L., H.K. and J.A. conceived the research. B.L. and R.A. performed the empirical analyses. B.L., R.A., Y.A., H.K. and J.A. discussed and interpreted the results and wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-025-02228-z.

Correspondence and requests for materials should be addressed to Haewoon Kwak or Jisun An.

Peer review information *Nature Human Behaviour* thanks Aida Mostafazadeh Davani, Hause Lin and Sam Zhang for their contribution to the peer review of this work. Peer reviewer reports are available.

nature portfolio

Corresponding author(s): Jisun An and Haewoon Kwak

Last updated by author(s): Apr 12, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection.
Data analysis	We developed custom code using Python 3.9.10. for data analysis, which is available at https://github.com/ByunghweeLee-IU/Belief-Embedding.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

We used a publicly accessible dataset from Debate.org (https://esdurmus.github.io/ddo.html). A processed version of this dataset, including pre-processed user-level debate records and the fine-tuned models used in our analyses, is available at https://github.com/ByunghweeLee-IU/Belief-Embedding.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	n/a
Reporting on race, ethnicity, or other socially relevant groupings	n/a
Population characteristics	n/a
Recruitment	n/a
Ethics oversight	n/a

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	This study is quantitative and utilizes data from an online debate forum (Debate.org) to examine the interplay between thousands of beliefs. A fine-tuned large language model (LLM) was employed to create an embedding space that represents human beliefs. This study explores how the belief space captures the interconnectedness and polarization of diverse beliefs, and tests whether this embedding space can be used to predict new beliefs of individuals.
Research sample	The research sample consists of user participation records from the Debate.org dataset, including a corpus of 78,376 debates (68,900 unique debate titles excluding duplicates) by 42,906 debaters from October 15, 2007, to September 19, 2018. While the dataset covers a broad range of social and political topics and captures diverse perspectives, it is primarily based on user-generated content from the United States, and thus may not fully represent global beliefs or cultural diversity. Nonetheless, we selected the Debate.org dataset because it offers a large-scale, structured collection of debate participation records and belief expressions of users across diverse social and political topics. Its format allows for direct mapping of users' positions on debates, making it well suited for modeling belief structures and studying patterns of polarization.
Sampling strategy	We utilized the entire Debate.org dataset (from October 15, 2007, to September 19, 2018).
Data collection	We utilized the existing publicly accessible Debate.org dataset (version 2, available at https://esdurmus.github.io/ddo.html), which includes user voting records on debates. Since the study used an existing public dataset, the researchers were not blinded to the hypothesis. However, the data were collected independently of this study and prior to hypothesis formulation.
Timing	We utilized the publicly accessible Debate.org dataset, which was downloaded on March 18, 2023, from https://esdurmus.github.io/ddo.html. The dataset covers a period from October 15, 2007, to September 19, 2018.
Data exclusions	Debate titles that could not be interpreted as belief statements were excluded from the dataset. For example, titles consisting of a single word or vague versus comparisons were filtered out using GPT-4 to ensure consistency in belief statement formulation. Of the 68,900 unique debate titles, 8,914 were removed during this filtering process, resulting in 59,986 debate titles retained for analysis.
Non-participation	No participants were involved in this study.
Randomization	Our study doesn't involve random-control design.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	n/a
Novel plant genotypes	n/a
Authentication	n/a