

The LLM Belief Atlas: Visualizing How Language-Model Stances Depend on Elicitation

Fan Huang*

Indiana University Bloomington

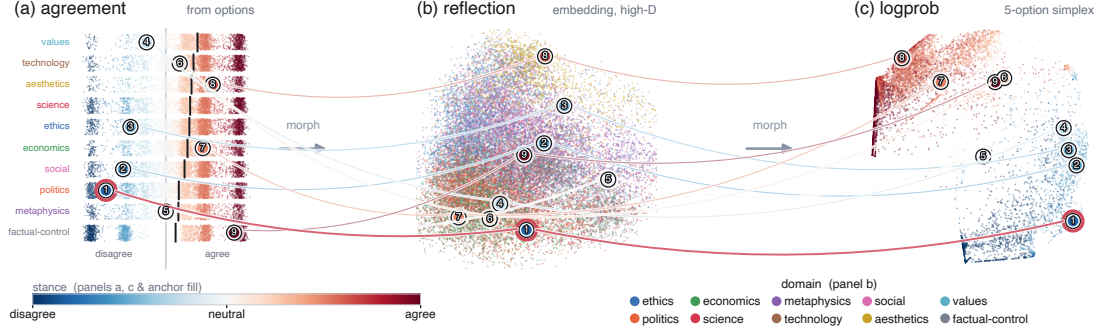


Figure 1: The *LLM Belief Atlas* co-registers three switchable readouts of the *same* elicited belief instances (four LLMs $\times$ 388 propositions $\times$ 49 steering conditions, 76,048 in total; a seeded 25,000-point subsample is drawn here) and morphs between them by preserved point identity: (a) *agreement* (a 5-level rating derived from the options, not separately declared), drawn as a constructed stance-by-domain layout because the projected agreement geometry is near-degenerate; (b) free-text *reflection* (a high-dimensional sentence embedding) forming topic clusters; and (c) option *logprob* (a 5-vector over the answer options, i.e. a point on the probability simplex) folding into a disagree-to-agree gradient. Each native readout is projected to the shared 2-D frame and summarized by a stance $s \in [-1, 1]$ (colorbar). Faint curves trace the nine numbered anchor beliefs; the bold rose exemplar is repositioned across three geometries.

ABSTRACT

What a language model appears to believe about a proposition is not a scalar but a signal that depends on how it is elicited and read: forcing a choice can fabricate a value, option-logprob scoring carries order bias, and sycophancy makes a stance a function of framing. We present the *LLM Belief Atlas*, a scalable in-browser WebGL instrument organized around a framework of *belief spaces*. The framework treats a model’s stance as a latent construct observed only through a four-dimensional measurement space, model, item, elicitation condition, and readout channel, and names seven subspaces, each bound to a facet of psychometric construct validity. The instrument co-registers three switchable belief readouts (Likert agreement, free-text reflection, and option logprobs) of the *same* elicitation instances and morphs between them at the million-point scale, while exposing uncertainty. On 76,048 elicited beliefs from four LLMs, framing conditions explain 34.2% of stance variance, against 46.8% for belief itself (and on par with it on the two natively-read models), but a within-model reliability measure does not predict cross-model reproducibility beyond noise.

Index Terms: Large language models, belief elicitation, measurement, construct validity, dimensionality reduction, scalable scatter-plot visualization, uncertainty visualization.

1 INTRODUCTION

Large language models increasingly mediate decisions about ethics, politics, and policy, so what a model appears to “believe” has become an object of measurement [16, 3, 17, 12, 7, 13]. Yet the

elicited signal is fragile: choice-forcing, option-order, paraphrase, social-context, and persuasion effects, together with sycophancy, each turn a stance into a function of how it is asked [15, 23, 18, 22, 19, 8, 10]. Eliciting beliefs across many models, items, framings, and channels therefore yields a structured object central to auditing alignment and political skew, yet rarely studied as such: prior belief studies [16, 3, 17] commit to a *single* channel and cannot separate a stance from the instrument that measured it. We contribute (i) a *framework of belief spaces* that organizes this object through psychometric construct validity, (ii) the *LLM Belief Atlas*, a scalable WebGL instrument whose *morph* keeps each answer’s identity as it moves between the three readouts, so disagreement between channels becomes directly visible, and (iii) a preliminary measurement study on 76,048 elicited beliefs that tests the framework’s predictions.

2 A FRAMEWORK OF BELIEF SPACES

A model’s belief about a proposition is not something we can read off directly. We can only estimate it through instruments that are themselves fallible, which is what construct validity [2, 1] was built to handle. Every observation is one point in a four-dimensional space: which model, which proposition, which *elicitation condition* (how the question was framed), and which *readout channel* (how the answer was scored). Format is a channel choice rather than a condition, so the two are crossed. The framework names seven subspaces, each varying one of those dimensions and holding the rest, and each bound to a facet of construct validity. Under this view the elicitation effects others log as caveats become measurement properties we can quantify.

3 THE BELIEF ATLAS INSTRUMENT

The atlas is a zero-build WebGL2 technique that co-registers three readouts of each instance, an *agreement* channel, a *reflection* em-

*e-mail: huangfan@acm.org

Table 1: One elicited belief: “Generous welfare programs reduce people’s incentive to seek employment”, read from four LLMs at the neutral default as Likert agreement (0–4, 0 = strongly disagree), option-logprob stance ($s \in [-1, 1]$), and uncertainty ($H \in [0, 1]$); the free-text reflection channel is shown in Fig. 1(b). The models disagree on the same proposition.

	Likert	s	H
Llama-3.3-70B	3	+0.50	0.03
Qwen-2.5-72B	2	0.00	0.72
DeepSeek-V3	2	0.00	0.72
GPT-4o-mini	1	−0.37	0.36

bedding of the free-form justification, and the option-logprob distribution, and renders interactively across the 10^6 – 10^7 regime [11, 14] (synthetic clouds on an NVIDIA A100, far above the 76,048 studied here: 455 fps at 10^6 points-mode, and 72–105 fps at 10^7). Each channel is projected and Procrustes-aligned onto an anchor over index-matched points; the index-correspondence *morph* then tweens between channels because point identity is preserved, so one point compares a belief under two readouts. A GPU kernel-density mode splats Gaussians and tone-maps density through a perceptually uniform colormap, so density reflects the data rather than overdraw [4]. Uncertainty (entropy) and inconsistency (Jensen–Shannon dispersion) are exported as explicit perceptual channels; projection distortion [21] is a known hazard the instrument does not yet quantify.

4 PRELIMINARY FINDINGS

We elicited 76,048 beliefs from four LLMs (Llama-3.3-70B, Qwen-2.5-72B, DeepSeek-V3, GPT-4o-mini) over 388 propositions under 49 steering conditions, a fully crossed model $\times$ item $\times$ condition design; uncertainty is read from native logprobs for 37,154 cells and estimated from at most three samples, re-elicited in October 2026, for the other 38,894; the steering conditions span paraphrase, negation, option order and labeling, persona, and user pressure. The same model’s stance also swings under reframing alone: Llama moves from +0.50 here to −0.47 under a Sweden persona. A generalizability-theory variance decomposition [20] partitions the elicited stance into belief structure, the framing condition, and residual noise, which explain 46.8%, 34.2%, and 19.0% of stance variance across all four models (that split collapses seven components under a rule we chose, and other defensible rules move the headline); on the two natively-read models ($n = 38,024$) the condition share (40.8%) is on par with belief structure (42.6%; residual 16.6%). The framework makes two falsifiable predictions. The first, condition $\times$ channel independence, is not testable here, because the two readouts spanning conditions are not independent: one is a rounding of the other. The second—that within-model reliability gates cross-model disagreement—is not supported: more inconsistent items have less reproducible cross-model ordering (Spearman $\rho = -0.75$, $n = 388$), but noise with the same per-item spread does nearly as well ($\rho = -0.73 \pm 0.02$). On a separate default-condition subset ($n = 1,552$), a multitrait–multimethod decomposition [1] attributes 52%, 17% and 7% of the reflection, logprob and agreement channels’ variance to method. Across the full corpus, sycophancy is asymmetric (−0.95 under disagree-pressure vs. +0.25 under agree-pressure).

5 CONCLUSION AND FUTURE WORK

The *LLM Belief Atlas* morphs three co-registered belief readouts at the million-point scale on a construct-validity framework. Its wider claim: a model’s stance is not a single number but a *field* that shifts with how a belief is elicited. Making that field auditable in one open instrument turns “what does this AI believe?” into

a reproducible measurement question for auditing, synthetic surveys [5], and honesty evaluation, each contingent on human validation not yet run [9, 6]. The perception study is specified and built, but piloted only with one annotator over ten trials; a scaled multi-participant run and probe/behavioral channels [7, 10] are the immediate next steps.

Three cautions. The native-logprob contrast rests on two models, so readout and model identity are confounded there. Option wording, ordering, tokenization and calibration are uncontrolled, and all four shape an option-probability reading. And we have not compared this instrument against a simpler visualization, so whether its complexity earns its keep is untested.

AI usage. The language models studied here are what we measured, not tools used to write it. An AI agent was applied to make grammar checks, and AI coding assistants supported software development; the author designed the study and verified all data, analyses, statistics, and claims. *Data.* No human-subject or personal data; persona answers are not opinion data. *Correction.* An answer-scoring defect in the poster version is fixed here.

REFERENCES

- [1] D. T. Campbell and D. W. Fiske. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2):81–105, 1959. 1, 2
- [2] L. J. Cronbach and P. E. Meehl. Construct validity in psychological tests. *Psychological Bulletin*, 52(4):281–302, 1955. 1
- [3] E. Durmus et al. Towards measuring the representation of subjective global opinions in language models. In *COLM*, 2024. 1
- [4] J. Heer. Fast & accurate Gaussian kernel density estimation. In *IEEE VIS*, pp. 11–15, 2021. 2
- [5] F. Huang, M. Kim, C. T. Diggans, and F. Radicchi. Evaluating LLM-generated preference distributions. *arXiv preprint arXiv:2610.01000*, 2026. 2
- [6] F. Huang, H. Kwak, and J. An. Is ChatGPT better than human annotators? Potential and limitations of ChatGPT in explaining implicit hate speech. In *WWW Companion*, pp. 294–297, 2023. 2
- [7] F. Huang, H. Kwak, and J. An. Understanding moral reasoning trajectories in large language models: Toward probing-based explainability. In *Findings of EMNLP*, 2026. To appear. arXiv:2603.16017. 1, 2
- [8] F. Huang, H. Kwak, and J. An. Vulnerability of LLMs’ stated belief? LLMs belief resistance check through strategic persuasive conversation interventions. In *Findings of ACL*, pp. 41759–41794, 2026. 1
- [9] F. Huang, H. Kwak, K. Park, and J. An. ChatGPT rates natural language explanation quality like humans: But on which scales? In *LREC-COLING*, pp. 3111–3132, 2024. 2
- [10] F. Huang, S. Zhang, H. Kwak, and J. An. CogBias: Measuring and mitigating cognitive bias in large language models. *arXiv preprint arXiv:2604.01366*, 2026. 1, 2
- [11] F. Lekschas. Regl-scatterplot: A scalable interactive JavaScript-based scatter plot library. *J. Open Source Softw.*, 8(84):5275, 2023. 2
- [12] W. Qi, F. Huang, J. An, and H. Kwak. A cross-cultural comparison of LLM-based public opinion simulation: Evaluating Chinese and U.S. models on diverse societies. In *IEEE ICDM Workshops (ICDMW)*, pp. 147–156, 2025. 1
- [13] W. Qi, F. Huang, R. Muralidharan, J. An, and H. Kwak. XChoice: Explainable evaluation of AI-human alignment in LLM-based constrained choice decision making. In *Findings of AACL-IJCNLP*, 2026. To appear. arXiv:2601.11286. 1
- [14] D. Ren, F. Hohman, H. Lin, and D. Moritz. Embedding atlas: Low-friction, interactive embedding visualization. In *IEEE VIS*, pp. 191–195, 2025. 2
- [15] P. Röttger et al. Political compass or spinning arrow? Towards more meaningful evaluations for values and opinions in large language models. In *ACL*, pp. 15295–15311, 2024. 1
- [16] S. Santurkar, E. Durmus, F. Ladhak, C. Lee, P. Liang, and T. Hashimoto. Whose opinions do language models reflect? In *ICML*, pp. 29971–30004. PMLR, 2023. 1
- [17] N. Scherrer, C. Shi, A. Feder, and D. M. Blei. Evaluating the moral beliefs encoded in LLMs. *NeurIPS*, 36:51778–51809, 2023. 1

- [18] J. Shao. First token probabilities are unreliable indicators for LLM knowledge. Technical Report UCB/EECS-2024-114, EECS Department, University of California, Berkeley, May 2024. [1](#)
- [19] M. Sharma et al. Towards understanding sycophancy in language models. In *ICLR*, 2024. [1](#)
- [20] R. J. Shavelson, N. M. Webb, and G. L. Rowley. Generalizability theory. *American Psychologist*, 44(6):922–932, 1989. [2](#)
- [21] M. Wattenberg, F. Viégas, and I. Johnson. How to use t-SNE effectively. *Distill*, 1(10):e2, 2016. [2](#)
- [22] L. Yin and F. Huang. DIF: A framework for benchmarking and verifying implicit bias in LLMs. *arXiv preprint arXiv:2505.10013*, 2025. [1](#)
- [23] C. Zheng, H. Zhou, F. Meng, J. Zhou, and M. Huang. Large language models are not robust multiple choice selectors. In *ICLR*, 2024. [1](#)