Emulate or Estimate? The Divergent Strengths of Base and Post-Trained Language Models for Opinion Simulation

Anonymous ACL submission

Abstract

Large language models are increasingly used to simulate human opinions, but prior work reports conflicting results: some studies find promising alignment with human survey data, while others find persona collapse and weak demographic sensitivity. We show that much of this conflict stems from conflating two distinct tasks. We call the first task emulation, in which models generate individual responses that aggregate into a population distribution. We call the second task estimation, in which models directly predict the population distribution. Evaluating six matched base and post-trained models on the Pew American Trends Panel, we find that base models are stronger emulators: they produce response distributions closer to human ground truth and better preserve demographic structure. Post-trained models are stronger estimators, producing more accurate distributional predictions when asked directly. We propose that model selection for human simulation should be guided by whether the task requires generating text or predicting distributions.

1 Introduction

Large Language Models (LLMs), trained on large corpora of human text, are well-positioned to act as simulators of human opinion (Argyle et al. (2023), Durmus et al. (2024)). High-fidelity population modelling would enable applications across policy testing, market research, and social science, motivating a growing body of work on LLM-based opinion simulation (Cao et al., 2025). However, prior results are conflicted: some works report promising distributional alignment with human survey data using base models (Suh et al. (2025), Moon et al. (2024)), while others document persona collapse (Li et al., 2025), demographic insensitivity (Sun et al., 2025), and sensitivity to methodological choices (Dominguez-Olmedo et al. (2024), Wang et al. (2024)). Studies reporting negative results evaluate primarily on post-trained models.

We propose that much of this disagreement reflects a conflation of two distinct simulation tasks: *emulation,* in which the model generates responses as an individual participant and the population distribution emerges from aggregation, and *estimation,* in which the model explicitly predicts the population's response distribution. We evaluate three matched base/post-trained pairs under both paradigms on the Pew American Trends Panel.

Our results demonstrate that base models produce emulated distributions consistently closer to ground truth, tracking demographic structure more faithfully. Post-trained models excel at estimation, producing well-calibrated distributional predictions. These findings clarify model selection for human simulation: base models are better suited for tasks requiring generated text, while post-trained models are better suited for predicting distributions directly.

2 Related Work

LLMs as human simulators. Language models have been increasingly explored as substrates for simulating human behaviour, spanning survey response simulation (Argyle et al. (2023), Cao et al. (2025), Park et al. (2026)), interactive social agents (Park et al. (2023), Vezhnevets et al. (2023)), and synthetic population modelling (Sun et al. (2024)). These efforts share the premise that pretraining corpora encode sufficient human diversity to support behavioural simulation at the population level. Following prior work (Santurkar et al. (2023), Suh et al. (2025)), we evaluate on survey opinion distributions as a controlled setting with ground truth.

Negative results and methodological concerns. A parallel literature has documented significant limitations. Recent works find that post-trained models exhibit systematic persona collapse, producing ho-

mogeneous outputs across prompted identities (Li et al. (2025), Xiao et al. (2026), Qin et al. (2026)). Sun et al. (2025) concluded that sociodemographic prompting does not reliably shift model outputs toward target demographics. Methodologically, Tjuatja et al. (2024), Dominguez-Olmedo et al. (2024), and Wang et al. (2024) demonstrated that model responses are sensitive to opinion ordering and label assignment. Hullman et al. (2026) argue that the field remains bottlenecked on prediction quality. These studies evaluate primarily or entirely on post-trained models.

Base models and distributional measurement. Several independent findings motivate the present work. Moon et al. (2024), Suh et al. (2025), and Cao et al. (2025) explicitly chose base models, citing opinion-skew in post-trained variants. Santurkar et al. (2023) observed that base models show closer distributional alignment with human survey data. Separately, Meister et al. (2025) showed that models can describe opinion distributions more accurately than they can produce them through sampling. We synthesize these findings by introducing the emulation-estimation distinction, systematically comparing matched base and post-trained pairs, and proposing open-response emulation as a positional-bias-free evaluation method.

3 Methodology

We study the problem of recovering human opinion distributions from language models: given a survey question and an optional demographic condition, produce a distribution over answer options that matches the empirical response distribution of the target population. Survey opinion data provides a controlled evaluation setting with ground-truth distributions across known demographic subgroups, making it a natural testbed for population simulation capabilities. We distinguish two paradigms for this task, and evaluate each on matched base and post-trained model pairs. See Figure 2 in Appendix A.1 for the evaluation flowchart.

3.1 Emulation Paradigm

Emulation recovers a population distribution by aggregating individual model responses. We implement this via open-response generation: the model produces free-text interview responses without access to multiple-choice options, and an LLM judge maps each response to an answer category (Cohen's $\kappa = 0.66$ with expert annotator, full validation in Appendix A.2.1). For base models, generation is structured as an interviewer-participant dialogue, with demographic conditioning introduced via a preceding dialogue turn. Post-trained models are given a system instruction to simulate a participant of a given demographic. We compare emulated distributions against human ground truth and a uniform reference. Prompt scaffolds for emulation can be found in Appendix A.2.

We choose the open-response sampling method over first-token probability extraction, in which the model's logit distribution over answer tokens serves as a direct distributional representation. While this has been widely used across prior works (Santurkar et al. (2023), Cao et al. (2025), Suh et al. (2025)), it is confounded by positional bias (Wang et al. (2024), Tjuatja et al. (2024)). Moreover, if the goal of emulation is to produce text that can be validated, first-token extraction offers no interpretable output. We adopt open-response emulation to eliminate the positional confounder, and present the first-token results — which corroborate our main findings — as Appendix A.6.

3.2 Estimation Paradigm

In the estimation paradigm, the model directly predicts the population's distribution for a given survey question, following the verbalized distribution approach introduced by Meister et al. (2025). The model receives the question with multiple-choice options and is prompted to return a JSON object representing the estimated probability distribution over answers. Prompt scaffolds for estimation can be found in Appendix A.3.

Post-trained models produce stable outputs under greedy decoding, so a single sample is taken. Base model completions are more variable, so we sample multiple completions and average the resulting distributions. We avoid providing a few-shot example format so as not to influence the prior. Both model types are prompted with appropriate formats to estimate the response distribution of a given demographic group.

3.3 Metrics

We evaluate distributional fidelity using two complementary metrics. Total Variation Distance (TVD) measures assumption-free distributional divergence. Wasserstein distance additionally respects the ordinal structure of answer options, penalizing errors proportionally to their distance from the correct response. Both metrics are standard in

the distributional alignment literature (Santurkar et al. (2023), Moon et al. (2024), Meister et al. (2025)). Full definitions are in Appendix A.4.

4 Experiments

4.1 Experimental Setup

Our evaluation dataset consists of 59 four-option items from the Pew Research Center's American Trends Panel Wave 54 (Pew Research Center, 2020), which specifically elicits economic opinions. Refusal responses are dropped and human distributions renormalized. We apply seven demographic conditions: an unconditioned marginal and six demographics spanning political (Democrat, Republican), ideological (Very Liberal, Very Conservative), economic (Upper-income), and religious (Protestant) axes. We evaluate six open-weight models with matched base and post-trained variants: Qwen3-14B, Olmo-3-7B, and Olmo-3.1-32B (Yang et al. (2025), Olmo et al. (2026)). Post-trained models are sampled at temperature 1.5 in emulation to support output diversity. We also include Claude Opus 4.6 (Anthropic, 2026) as a frontier estimation reference.

4.2 Base Models are Emulators

Base models produce emulated distributions consistently closer to human ground truth than their post-trained counterparts across all six models and all seven evaluation conditions (Table 1). On Wasserstein distance, base models outperform their post-trained counterparts in every setting and beat the uniform reference in 20 of 21 model-condition comparisons, indicating that even when base models miss exact categorical probabilities, mass is placed near the correct ordered responses. On TVD, base models again outperform their post-trained counterparts in every setting, though they beat the uniform reference distribution in only 11 of 21 comparisons — a weaker margin than Wasserstein, where ordinal proximity to correct answers is rewarded. The per-condition comparison of Olmo-32B base and post-trained variants confirms this advantage is consistent across all six demographic conditions (Figure 1b). This gap is driven by mode collapse in post-trained outputs, which are 8.0-14.0× more similar at the bigram level than base model outputs (Table 4); elevated temperature does not resolve this collapse.

To test whether conditioned models track genuine demographic structure, we correlate model and human pairwise TVD distances among the six conditioned groups (15 pairs) via Spearman's ρ (Table 2). Base models consistently align with human demographic structure ($\rho = 0.61$-0.75, all $p < 0.02$) while compressing inter-group differences to approximately 70% of human magnitude. Post-trained models show weaker structural alignment ($\rho = 0.35$-0.59) and exaggerate inter-group differences by approximately 2×, consistent with stereotyped persona simulation (Li et al., 2025).

Table 1: Average error across the seven evaluation conditions. Lower is better.

	Estimation error		Emulation error	
Model	TVD	Wasserstein	TVD	Wasserstein
Qwen3-14B Base	0.236	0.492	**0.263**	**0.457**
Qwen3-14B	**0.213**	**0.422**	0.461	0.707
Olmo-3-1025-7B	**0.278**	0.605	**0.248**	**0.415**
Olmo-3-7B-Instruct	0.285	**0.603**	0.372	0.573
Olmo-3-1125-32B	0.223	0.460	**0.274**	**0.443**
Olmo-3.1-32B-Instruct	**0.185**	**0.347**	0.427	0.643
Claude Opus 4.6	**0.142**	**0.267**	—	—
uniform	0.271	0.582	—	—

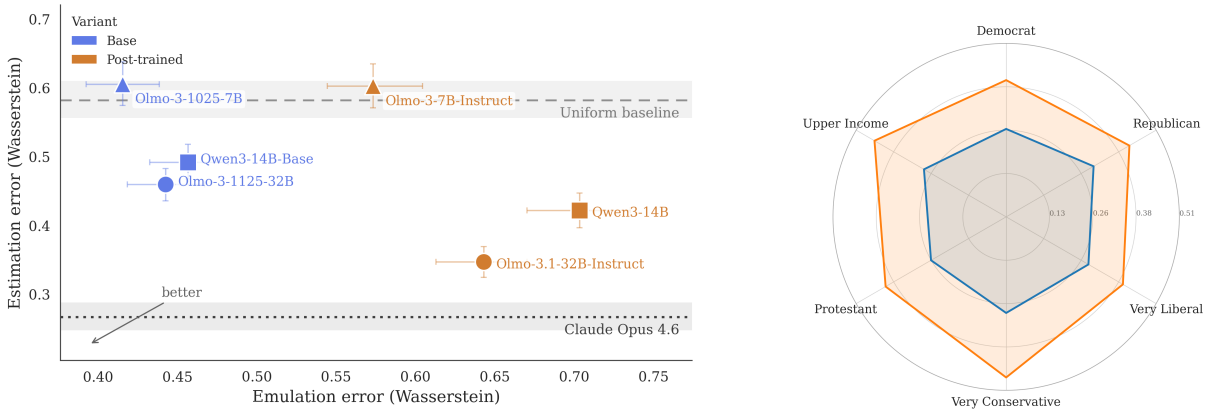
Table 2: Spearman correlations and mean ratio of inter-group distribution between open-response model and human pairwise distances.

Model	ρ	p	mean ratio
Qwen3-14B Base	0.754	0.001	0.688
Qwen3-14B	0.536	0.040	2.242
Olmo-3-1025-7B	0.682	0.005	0.664
Olmo-3-7B-Instruct	0.350	0.201	2.014
Olmo-3-1125-32B	0.607	0.016	0.704
Olmo-3.1-32B-Instruct	0.593	0.020	2.063

4.3 Post-trained Models are Estimators

Under the estimation paradigm, post-trained models produce distributions substantially closer to human ground truth than the uniform baseline across all conditions (Table 1). Estimation error decreases with model scale and is substantially lower in post-trained models than their base counterparts, with Claude Opus 4.6 achieving the strongest performance. These results are consistent with the finding of Meister et al. (2025) that distributional knowledge is accessible through explicit verbalization.

Base models also produce estimation results that match or exceed the uniform baseline in most configurations, indicating that distributional knowledge about human opinion is present prior to post-training.



(a) Emulation-estimation tradeoff across base and post-trained models. Each point represents mean Wasserstein distance against human ground-truth distributions, averaged across seven conditions. Lower is better on both axes. Error bars: 95% bootstrap CIs from question resampling (2,000 replicates). Baselines shown with uncertainty bands.

(b) Per-condition emulation error (TVD) for Olmo-3-1125-32B and Olmo-3.1-32B-Instruct. Smaller area indicates closer alignment with human distributions.

Figure 1: Summary of model error patterns. Base models are blue and post-trained models are orange.

5 Discussion

Our results suggest that opinion simulation should be evaluated according to the form of output required by the downstream task. If the goal is to estimate aggregate opinion distributions, post-trained models are the more appropriate substrate. If the goal is to generate text that aligns with a population's diversity — as would be required for synthetic respondents or interactive agents — base models are better aligned with the task, because the desired distribution must emerge from sampled outputs rather than from a direct verbal estimate. More broadly, evaluations should measure whether generated responses preserve both aggregate opinion distributions and demographic structure, rather than only whether a model can describe those distributions when prompted directly.

This distinction also clarifies why prior findings on LLM opinion simulation appear mixed. Evaluations based on explicit prediction and evaluations based on generated responses measure different capabilities. Post-training appears to improve distributional estimation while reducing the output diversity needed for faithful emulation. The Persona Selection Model (Marks et al., 2026) offers a theoretical account of this tradeoff. Post-training steers the base model's broad, chimeric distribution into a single Assistant persona optimized for helpfulness (Lu et al., 2026). When asked to emulate a demographic group, the post-trained model performs *second-order simulation* — an Assistant persona simulating a demographic persona — which compresses output diversity through the bottleneck of a single character. The Assistant persona produces well-calibrated estimates because the underlying LLM has learned that a helpful assistant, when asked a factual question, will draw on its available knowledge to provide a useful answer (Shanahan et al., 2023). Base models bypass this bottleneck: their outputs draw directly from the pre-trained distribution over personas, preserving more of the representative diversity needed for emulation at the cost of weaker metacognitive estimation.

6 Conclusion

We distinguish two paradigms for simulating human opinion distributions with language models: emulation and estimation. Post-trained models excel at estimation, producing well-calibrated distributional predictions when asked explicitly. Base models excel at emulation, generating response distributions that more faithfully track human demographic structure while compressing inter-group differences to approximately 70% of human magnitude, compared to the approximately 2× exaggeration observed in post-trained models. We propose open-response emulation as an evaluation paradigm free of the positional bias that confounds first-token extraction, and argue that the choice of simulation substrate should be guided by whether the downstream task requires distributional estimates or generated text.

Limitations

Whether the emulation-estimation tradeoff persists or collapses at frontier scale is unknown, as frontier base model variants are not released.

A second limitation concerns temporal alignment. The Pew survey data was collected in 2019, while the models were trained on corpora extending through 2024-2025. Human opinion distributions shift over time, and it is unclear whether models simulate the opinions of a particular era or encode a temporally blended representation. This confound applies to both base and post-trained models. Whether effective population simulation will require continually updated models, or whether temporal dynamics can be disentangled mechanistically, remains an open question requiring community-wide effort.

Finally, our evaluation is limited to survey opinion; extending the capabilities of base models to richer behavioural settings, including multi-turn dialogue and agent-based simulation, is a natural step toward the broader goal of faithful population modelling.

Ethical Considerations

Simulating demographic opinion distributions requires conditioning on broad categories, which risks reinforcing stereotypes if models collapse groups to caricatured responses. We adopt distributional evaluation, measuring how well models capture the full spread of opinion within groups rather than a single modal response. While we argue that base models are a more suitable substrate for text-generating simulation, we caution that any deployment as synthetic respondents should be validated against real population data before informing decisions. All human ground-truth data is drawn from the Pew Research Center's American Trends Panel, a consented and anonymized survey instrument.

References

Anthropic. 2026. Claude opus 4.6 system card. Technical report, Anthropic.

Lisa Argyle, Ethan Busby, Nancy Fulda, Joshua Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31:1–15.

Yong Cao, Haijiang Liu, Arnav Arora, Isabelle Augenstein, Paul Röttger, and Daniel Hershcovich. 2025. Specializing large language models to simulate survey response distributions for global populations. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3141–3154, Albuquerque, New Mexico. Association for Computational Linguistics.

Ricardo Dominguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dünner. 2024. Questioning the survey responses of large language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Esin Durmus, Karina Nguyen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2024. Towards measuring the representation of subjective global opinions in language models. *Preprint*, arXiv:2306.16388.

Jessica Hullman, David Broska, Huaman Sun, and Aaron Shaw. 2026. This human study did not involve human subjects: Validating llm simulations as behavioral evidence. *Preprint*, arXiv:2602.15785.

Ang Li, Haozhe Chen, Hongseok Namkoong, and Tianyi Peng. 2025. Llm generated persona is a promise with a catch. *Preprint*, arXiv:2503.16527.

Christina Lu, Jack Gallagher, Jonathan Michala, Kyle Fish, and Jack Lindsey. 2026. The assistant axis: Situating and stabilizing the default persona of language models. *Preprint*, arXiv:2601.10387.

Sam Marks, Jack Lindsey, and Christopher Olah. 2026. The Persona Selection Model: Why AI assistants might behave like humans. `https://alignment.anthropic.com/2026/psm/`.

Nicole Meister, Carlos Guestrin, and Tatsunori Hashimoto. 2025. Benchmarking distributional alignment of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 24–49, Albuquerque, New Mexico. Association for Computational Linguistics.

Suhong Moon, Marwa Abdulhai, Minwoo Kang, Joseph Suh, Widyadewi Soedarmadji, Eran Kohen Behar, and David M. Chan. 2024. Virtual personas for language models via an anthology of backstories. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19864–19897, Miami, Florida, USA. Association for Computational Linguistics.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish

Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, and 49 others. 2026. Olmo 3. *Preprint*, arXiv:2512.13961.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA. Association for Computing Machinery.

Joon Sung Park, Carolyn Q. Zou, Jonne Kamphorst, Niles Egan, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Percy Liang, Robb Willer, and Michael S. Bernstein. 2026. Llm agents grounded in self-reports enable general-purpose simulation of individuals. *Preprint*, arXiv:2411.10109.

Pew Research Center. 2020. American Trends Panel Wave 54. `https://www.pewresearch.org/dataset/american-trends-panel-wave-54/`. Fielded September 16–29, 2019. Topic: Economic inequality.

Xiaoyou Qin, Zhihong Li, and Xiaoxiao Cheng. 2026. Restoring heterogeneity in llm-based social simulation: An audience segmentation approach. *Preprint*, arXiv:2604.06663.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature*, 623(7987):493–498.

Joseph Suh, Erfan Jahanparast, Suhong Moon, Minwoo Kang, and Serina Chang. 2025. Language model fine-tuning on scaled survey data for predicting distributions of public opinions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21147–21170, Vienna, Austria. Association for Computational Linguistics.

Huaman Sun, Jiaxin Pei, Minje Choi, and David Jurgens. 2025. Sociodemographic prompting is not yet an effective approach for simulating subjective judgments with LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 845–854, Albuquerque, New Mexico. Association for Computational Linguistics.

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. *Preprint*, arXiv:2402.18144.

Lindia Tjuatja, Valerie Chen, Tongshuang Wu, Ameet Talwalkwar, and Graham Neubig. 2024. Do LLMs exhibit human-like response biases? a case study in survey design. *Transactions of the Association for Computational Linguistics*, 12:1011–1026.

Alexander Sasha Vezhnevets, John P. Agapiou, Avia Aharon, Ron Ziv, Jayd Matyas, Edgar A. Duéñez-Guzmán, William A. Cunningham, Simon Osindero, Danny Karmon, and Joel Z. Leibo. 2023. Generative agent-based modeling with actions grounded in physical, social, or digital space using concordia. *Preprint*, arXiv:2312.03664.

Hanna Wallach, Meera Desai, A Feder Cooper, Angelina Wang, Chad Atalla, Solon Barocas, Su Lin Blodgett, Alexandra Chouldechova, Emily Corvi, P Alex Dow, and 1 others. 2025. Position: Evaluating generative ai systems is a social science measurement challenge. *arXiv preprint arXiv:2502.00561*.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. 2024. "my answer is C": First-token probabilities do not match text answers in instruction-tuned language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7407–7416, Bangkok, Thailand. Association for Computational Linguistics.

Yunze Xiao, Vivienne J. Zhang, Chenghao Yang, Ningshan Ma, Weihao Xuan, and Jen tse Huang. 2026. The chameleon's limit: Investigating persona collapse and homogenization in large language models. *Preprint*, arXiv:2604.24698.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

A Appendix

A.1 Evaluation Setup

Figure 2 provides a high-level overview of the evaluation setup and the flow charts for the emulation and estimation sampling paradigms.

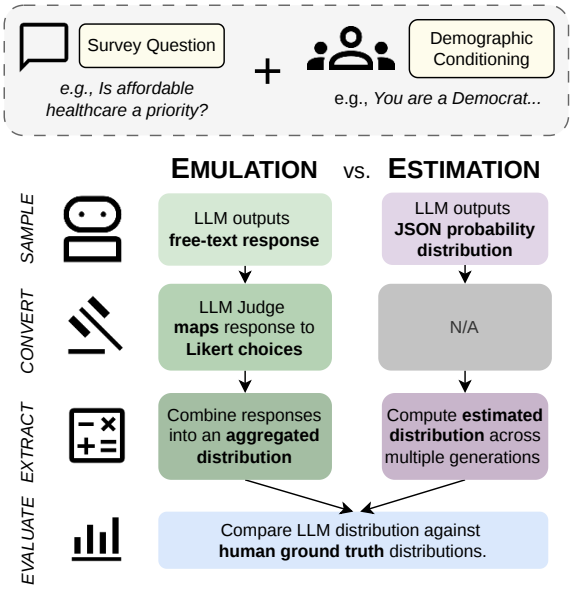


Figure 2: Overview of the evaluation method for **emulation** and **estimation**.

A.2 Emulation Setup

A.2.1 LLM Judge

We validate the LLM judge used in the emulation pipeline against human annotations on a sample of 200 response-category pairs. The human annotator is a member of the research team who had understanding of the research goals, in order to best capture construct validity (Wallach et al., 2025). On the full mapping task, which includes responses judged as unmappable, the judge achieves Cohen's $\kappa = 0.66$ with 74% raw agreement. Restricting to the 146 pairs where both the judge and human annotator assigned a letter category, agreement rises to $\kappa = 0.72$ (81% raw agreement).

Prompt

Read the survey response and map it to the best matching multiple-choice option. You are interpreting a language model's response; after the "answer" it may appear very messy (e.g. interview continuation). If the first part of a response is at all usable, i.e. looks like an answer continuation, map it, even if the text afterwords frames it in a nonstandard way. Return only a single option letter. If the response is unusable, return X. Interpret the answer at face value; do not try to reason about it or rationalize intentions. Do not infer a stronger or more socially desirable answer than what is actually stated.

Question: *[question text]*

Options: *[options]*

Response: *[response text]*

Figure 3: Judge prompt used to map free-form model responses to multiple-choice survey options.

A.2.2 Base Model Emulation Scaffolding

Results are computed across $N = 100$ stochastic generations per question.

Prompt

INTERVIEWER: How would you describe your household's financial situation?\nPARTICIPANT:

Figure 4: Prompt scaffold for unconditioned base-model emulation. The model is framed as a participant in an interview-style dialogue and generates a free-form survey response without access to answer options.

Prompt

INTERVIEWER: How would you describe yourself politically?\nPARTICIPANT: I am a Democrat.\nINTERVIEWER: How would you describe your household's financial situation?\nPARTICIPANT:

Figure 5: Prompt scaffold for demographically conditioned base-model emulation. Demographic conditioning is introduced through a preceding dialogue turn before the survey question is asked.

Table 3: Percent usable responses for base open-response models across the seven W54 evaluation settings. Each entry reports the average percentage of generations that the judge successfully mapped to one of the survey answer options, averaged over questions within that setting. Higher values indicate that a larger share of free-form generations could be used in evaluation.

Model	Marginal	Democrat	Republican	Very Lib.	Very Cons.	Protestant	Upper Inc.	Mean
Qwen3-14B Base	80	84	82	84	83	82	83	82
Olmo-3-1025-7B	79	81	78	82	78	75	81	79
Olmo-3-1125-32B	82	87	86	90	86	86	88	87

A.2.3 Post-trained Model Emulation Scaffolding

Similar to the base model setup, results are computed across $N = 100$ stochastic generations per question.

Prompt

`<system>`You are simulating a participant's answer to a survey question. Generate responses that reflect the full range of American opinion. Return only the participant's answer in plain text, with no markdown or explanation.

`<user>`Survey question: How would you describe your household's financial situation?

Figure 6: Prompt scaffold for unconditioned post-trained model emulation. The system instruction frames the model as a survey participant and requests a plain-text response to the survey question.

Prompt

`<system>`You are simulating a participant's answer to a survey question. Generate the response that a person with the following attribute would give. Attribute: <attribute>. Return only the participant's answer in plain text, with no markdown or explanation.

`<user>`Survey question: How would you describe your household's financial situation?

Figure 7: Prompt scaffold for demographically conditioned post-trained model emulation. The system instruction specifies a target demographic attribute and asks the model to generate a corresponding participant response.

A.2.4 Self-similarity Comparison

Table 4: Average pairwise Jaccard similarity for open-response generations, averaged across 59 questions and 7 W54 conditions. Responses are normalized and truncated to the first 25 words.

Model	Unigram	Bigram
Qwen3-14B Base	0.076	0.017
Qwen3-14B	0.368	0.238
Olmo-3-1025-7B	0.089	0.014
OLMo-3-7B-Instruct	0.243	0.142
OLMo-3-1125-32B	0.087	0.020
OLMo-3.1-32B-Instruct	0.274	0.159

A.3 Estimation Setup

A.3.1 Base Model Estimation Scaffolding

Prompt

The following is the estimated distribution of <population_text> to the question: "<question_text>"\nOptions: <options>\nEstimated distribution: {{"A":

Figure 8: Prompt scaffold for base-model estimation. The model directly completes a partially specified probability distribution over survey response options.

A.3.2 Post-trained Model Estimation Scaffolding

Prompt

<system>You are estimating the probability distribution of human answers to a multiple-choice survey question. Return exactly one JSON object and nothing else. Use every option letter exactly once as a key. Each value must be an integer percentage. The percentages must sum to 100. Do not include explanation, markdown, code fences, comments, or any keys other than the option letters.

<user>Target demographic: estimate the distribution for respondents matching these filters, not for the general population: <target_demographic>

Question: <question_text>

Options: <options>

Return a JSON object with integer percentages summing to 100.

Figure 9: Prompt scaffold for post-trained model estimation. The model is instructed to explicitly estimate the probability distribution of survey responses and return a JSON object with integer percentages summing to 100.

A.4 Metric Definitions

Let Q be the set of completed questions included in evaluation, and let $N = |Q|$. For each question $q \in Q$, let:

- H_q be the human response distribution over the options for question q
- M_q be the model response distribution over the options for question q
- $H_q(i)$ and $M_q(i)$ be the mass assigned to option i by H_q and M_q respectively
- $\mathcal{O}_q = \{1, \ldots, K_q\}$ be the option set for question q

A.4.1 Total Variation Distance

For each question q,

$$\mathrm{TVD}(H_q, M_q) = \frac{1}{2} \sum_{i \in \mathcal{O}_q} |H_q(i) - M_q(i)| \,.$$

The pipeline-level score is

$$\mathrm{MeanTVD} = \frac{1}{N} \sum_{q \in Q} \mathrm{TVD}(H_q, M_q).$$

A.4.2 Wasserstein Distance

We compute the Wasserstein-1 distance (earth mover's distance) over the ordered answer options per question from cumulative mass differences:

$$W_1(H_q, M_q) = \sum_{i=1}^{K_q - 1} \left| \sum_{j=1}^{i} \big(H_q(j) - M_q(j)\big) \right| .$$

The pipeline-level score is

$$\mathrm{MeanWasserstein} = \frac{1}{N} \sum_{q \in Q} W_1(H_q, M_q).$$

A.5 Per-Condition Breakdowns

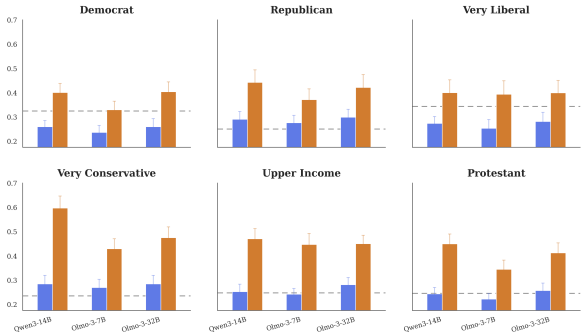


Figure 10: Per-condition emulation error (TVD). Lower is better. Base models have a consistent advantage over post-trained models.

See Table 5 and Table 6 for full per-condition emulation and estimation results respectively.

Table 5: Emulation per condition

Model	Marginal		Democrat		Republican		Very Lib.		Very Cons.		Protestant		Upper Inc.	
	TVD	Wasserstein	TVD	Wasserstein	TVD	Wasserstein	TVD	Wasserstein	TVD	Wasserstein	TVD	Wasserstein	TVD	Wasserstein
Qwen3-14B Base	0.237	0.398	0.260	0.469	0.290	0.490	0.273	0.492	0.284	0.503	0.243	0.417	0.252	0.427
Qwen3-14B	0.447	0.648	0.400	0.564	0.442	0.688	0.400	0.611	0.597	1.021	0.450	0.658	0.470	0.734
Olmo-3-1025-7B	0.236	0.386	0.237	0.402	0.276	0.453	0.253	0.443	0.270	0.454	0.222	0.372	0.242	0.397
Olmo-3-7B-Instruct	0.290	0.426	0.329	0.492	0.371	0.623	0.393	0.636	0.430	0.656	0.344	0.496	0.447	0.683
Olmo-3-1125-32B	0.253	0.389	0.260	0.427	0.299	0.483	0.281	0.488	0.284	0.463	0.257	0.408	0.281	0.441
Olmo-3.1-32B-Instruct	0.429	0.647	0.404	0.592	0.421	0.640	0.399	0.611	0.475	0.733	0.412	0.611	0.450	0.667
uniform reference distribution	0.251	0.547	0.325	0.712	0.250	0.503	0.343	0.768	0.235	0.498	0.245	0.525	0.247	0.520

Table 6: Estimation per condition

Model	Marginal		Democrat		Republican		Very Lib.		Very Cons.		Protestant		Upper Inc.	
	TVD	Wasserstein	TVD	Wasserstein	TVD	Wasserstein	TVD	Wasserstein	TVD	Wasserstein	TVD	Wasserstein	TVD	Wasserstein
Qwen3-14B Base	0.194	0.419	0.282	0.591	0.214	0.426	0.300	0.626	0.237	0.493	0.223	0.487	0.203	0.403
Qwen3-14B	0.193	0.407	0.208	0.418	0.227	0.446	0.223	0.430	0.258	0.482	0.198	0.402	0.187	0.370
Olmo-3-1025-7B	0.235	0.529	0.327	0.720	0.250	0.520	0.365	0.795	0.268	0.582	0.241	0.535	0.258	0.556
Olmo-3-7B-Instruct	0.175	0.351	0.318	0.698	0.256	0.523	0.387	0.866	0.370	0.760	0.255	0.538	0.236	0.483
Olmo-3-1125-32B	0.206	0.432	0.260	0.555	0.197	0.379	0.273	0.578	0.210	0.419	0.200	0.422	0.212	0.432
Olmo-3.1-32B-Instruct	0.170	0.335	0.191	0.376	0.181	0.333	0.216	0.401	0.215	0.370	0.153	0.293	0.169	0.321
Claude Opus 4.6	0.130	0.245	0.124	0.224	0.164	0.311	0.119	0.195	0.174	0.342	0.141	0.286	0.143	0.266
uniform baseline	0.251	0.547	0.325	0.712	0.250	0.503	0.343	0.768	0.235	0.498	0.245	0.525	0.247	0.520

A.6 First-token Extraction

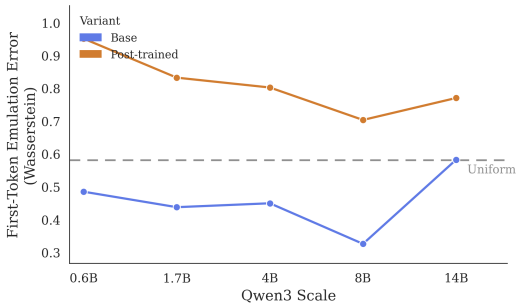


Figure 11: First-token emulation error (Wasserstein distance, averaged across 7 evaluation conditions) for Qwen3 base and post-trained models across 8B to 14B scale. Base models outperform their post-trained counterparts at every scale, with the strongest performance at 8B. The 14B base model degrades to near-uniform levels, motivating our adoption of open-response emulation for the main evaluation.

Scaffolds were selected to maximize probability mass over valid answer tokens, independent of distributional fit to human data. The base model prompt ends with "My choice is Letter:" to constrain generation to answer tokens. The post-trained model uses a system prompt framing the model as a survey participant.

First-token extraction shows a consistent base advantage across scales (Figure 11), with the strongest performance at 8B parameters. Performance degrades at 14B, potentially due to stronger positional priors. Models at this scale show evidence of positional-bias confounding, motivating our adoption of open-response emulation for the main evaluation. The directional finding that base models outperform instruction-tuned models is consistent with the open-response emulation results reported in the main paper.