

An Analysis of Large Language Models for Simulating User Responses in Surveys

Ziyun Yu¹², Yiru Zhou¹, Chen Zhao¹², Hongyi Wen¹²

¹Center for Data Science, NYU Shanghai ²New York University

zy2478@nyu.edu, yz8297@nyu.edu, cz1285@nyu.edu, hw3242@nyu.edu

1

Abstract

Using Large Language Models (LLMs) to simulate user opinions has received growing attention. Yet LLMs, especially trained with reinforcement learning from human feedback (RLHF), are known to exhibit biases toward dominant viewpoints, raising concerns about their ability to represent users from diverse demographic and cultural backgrounds. In this work, we examine the extent to which LLMs can simulate human responses to cross-domain survey questions through direct prompting and chain-of-thought (CoT) prompting. We further propose a claim diversification method (CLAIMSIM), which elicits viewpoints from LLM parametric knowledge as contextual input. Experiments on the survey question answering task indicate that, while CLAIMSIM produces more diverse responses, both approaches struggle to accurately simulate users. Further analysis reveals two key limitations: (1) LLMs tend to maintain fixed viewpoints across varying demographic features, and generate single-perspective claims; and (2) when presented with conflicting claims, LLMs struggle to reason over nuanced differences among demographic features, limiting their ability to adapt responses to specific user profiles.²

1 Introduction

The development of large language models (LLMs) has enabled simulating human behavior and replicating individual decision-making processes (Park et al., 2024; Binz and Schulz, 2024; Aher et al., 2023). For example, LLMs have been adopted to design large-scale market surveys and attempt to simulate responses across diverse demographic groups, providing a cost-effective alternative to traditional survey methods (Brand et al., 2024).

¹Corresponding author: Hongyi Wen <hw3242@nyu.edu>

²Our code and data are available at <https://github.com/Ziyun-Yu/ClaimSim>.

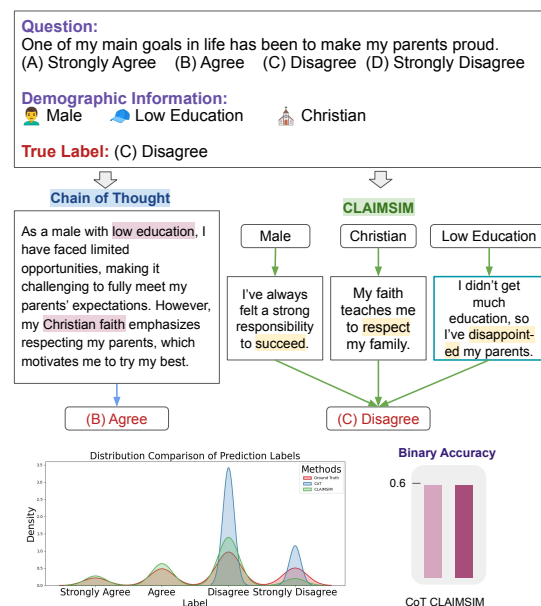


Figure 1: Top: A survey question answering example where LLMs are instructed to simulate individual user responses over diverse demographic profiles. Middle: We study two LLM-based approaches on this task, CoT and CLAIMSIM. Bottom: CLAIMSIM produces more diverse answers, while both approaches struggle to simulate users accurately (slightly above random).

Despite their strong potential, LLMs, particularly those trained with reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022), are known to exhibit significant bias issues (Schramowski et al., 2022; Messeri and Crockett, 2024; Hu et al., 2024). This is especially pronounced when interacting with underrepresented demographic or cultural groups (Wang et al., 2025; Santurkar et al., 2023). For instance, Wang et al. (2024) argues that LLM responses across different languages tend to align more closely with English-speaking cultural norms.

In this study, we systematically evaluate to what extent LLMs simulate human responses to survey questions, particularly in datasets like the World Values Survey (World Values Survey, 2020) where

respondents exhibit diverse and complex demographic profiles. As an example in Figure 1, LLMs are instructed to simulate a Christian male user with a lower level of education, and to answer a gender related survey question.

We compare three categories of LLM-based approaches. First, we use direct prompting (Beck et al., 2024) as a baseline to evaluate the models’ ability to do user simulations. Then, we incorporate chain-of-thought prompting (CoT) (Wei et al., 2022) to elicit LLMs’ capability of reasoning over user demographic features and simulating the answers accordingly. We further hypothesize that eliciting multiple viewpoints from LLM parametric knowledge as context leads to more comprehensive reasoning over diverse demographic features. Therefore, we present a two-step pipeline (CLAIMSIM): **(1) Claims Generator:** for each demographic feature, the LLM generates diverse claims and summarizes them, with instructions to explicitly highlight consistencies and contradictions; **(2) Answer Generator:** the summaries across all demographic features are used as additional context for the final answer simulation.

We experiment on randomly selected individuals and their survey answers drawn from the three domains: gender, politics, and religion. The results indicate that compared with DIRECT PROMPTING and CoT, CLAIMSIM provides more diverse answer distributions. By eliciting multiple claims, it encourages LLMs to consider a broader range of perspectives, resulting in more balanced responses. However, both approaches struggle to accurately simulate users. Our fine-grained analysis reveals that LLMs fail to reason over conflicting claims; and more importantly, for 50% of the survey questions, the claims generated by CLAIMSIM reflect a single-viewed opinion, regardless of demographic variation. Our findings reveal a fundamental limitation in LLMs’ capability of simulating user behavior: RLHF can lead to the entrenchment of certain viewpoints, therefore resistant to change across different contexts.

2 Method

This section first introduces the task formulation of survey question answering (§2.1), followed by two approaches, chain-of-thought prompting (§2.3), and CLAIMSIM (§2.4).

2.1 Task Formulation

We formulate the task of survey question answering as follows: The input consists of a tuple of (q, A, D) , where q refers to a multiple-choice question about opinions, A refers to the answer candidates, and D refers to a target individual’s basic demographic profile, which includes attributes such as sex, birth decade, and religion, among others. The objective is to instruct an LLM to simulate the perspective of the individual with the provided demographic background:

$$\phi_{LLM}(q, A, D) \Rightarrow a,$$

where $a \in A$ is the most probable answer selected by an LLM.

2.2 Direct Prompting

Our first approach uses direct prompting (Beck et al., 2024) to guide LLMs in answering survey questions. Given a question q and a set of demographic features D , an LLM is prompted to directly generate an answer a .

2.3 Chain-of-thought Prompting

We then use chain-of-thought prompting (Wei et al., 2022) to guide an LLM to first articulate its reasoning process by considering and integrating input features, and then to generate an answer a based on that reasoning step by step.

2.4 CLAIMSIM: Simulating Users with Diverse Claims Generation

We hypothesize that generating claims elicited from LLM parametric knowledge for each individual demographic feature as additional context could help mitigate model bias, as variations across features often lead to conflicting opinions. To implement this idea, we prompt LLMs several times to generate a set of representative claims $C_i = \{c_1, c_2, \dots, c_n\}$ for each demographic feature D_i based on query q :

$$\phi_{LLM}(q, D_i) \Rightarrow C_i,$$

The LLMs are then instructed to summarize these claims C_i into a single output S_i :

$$\phi_{LLM}(C_i) \Rightarrow S_i,$$

Our final answer prediction a is grounded on the input query q , demographic information D , and aggregated claim summaries $S = \{S_1, S_2, \dots, S_k\}$:

$$\phi_{LLM}(q, A, D, S) \Rightarrow a.$$

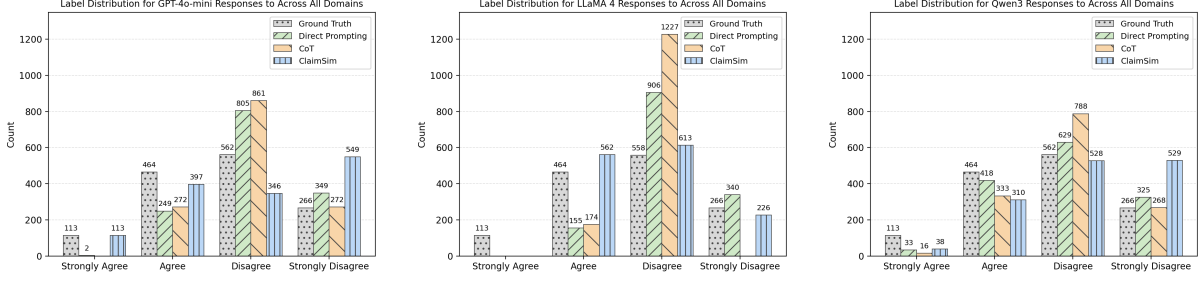


Figure 2: Comparison of answer distributions averaged across domains for DIRECT PROMPTING, CoT, and CLAIMSIM (left to right with GPT-4O-MINI, LLAMA 4, and QWEN 3). CLAIMSIM leads to more diverse answer distributions.

Diverse Claims Generation. For each demographic feature, we first prompt the LLM to generate claims C_i in response to the corresponding survey question q . To ensure diverse perspectives, we sample five separate responses, each producing one claim. We then instruct the LLM to summarize these claims into a single output S_i , explicitly highlighting both consistencies and contradictions within responses.

Answer Prediction. With several claim summaries derived from different demographic features, we then prompt the LLM to answer the corresponding survey question. Using chain-of-thought prompting, we guide the model to consider both the demographic information and the potentially conflicting claims, to more accurately simulate a user with a specific demographic background.

3 Experiments

This section first presents the dataset and experiment settings (§3.1 & 3.2). Then we provide detailed results (§3.3) and analysis (§3.4) for our survey question answering task.

3.1 Dataset

Our dataset is derived from the World Values Survey (WVS), a comprehensive global database that captures a broad range of demographic attributes and value-based attitudes across diverse populations (World Values Survey, 2020). For our experiments, we focus on three domains with the presence of strongest opinions—**gender, politics, and religion**—as a representative subset, covering a total of 16 questions. We randomly sample 100 individuals from the full dataset, ensuring substantial demographic diversity within the selection (details on sample counts are provided in the Appendix A).

3.2 Setups

Models We evaluate several LLMs, including the proprietary model GPT-4O-MINI (OpenAI, 2025), and open-source models LLAMA4 17B (Meta AI, 2025) and QWEN3 235B-A22B (Qwen Team, 2025). For the proprietary model, experiments are conducted via the OpenAI API. The open-source models are accessed through Together AI’s API service. For all experiments, the temperature parameter is set to 0.7 whenever applicable, as this is a common and widely used value in OpenAI’s documentation (OpenAI, 2024).

Metrics Our primary evaluation metric is Answer Accuracy (**Acc**), defined as the exact match with the gold answer option. We also report Binary Answer Accuracy (**B-Acc**), which maps nuanced labels (e.g., strongly agree and agree) into two attitudinal categories: agree and disagree. Additionally, we measure **Diversity** by analyzing the distribution of predicted answers with a histogram. To further qualitatively validate the differences between distributions produced by different methods, we compute the Wasserstein Distance, a metric that quantifies the shift between probability distributions (Villani et al., 2008).

3.3 Main Results

DIRECT PROMPTING and CoT do not have significant differences. As shown in Table 1 and Table 2, DIRECT PROMPTING and CoT achieve comparable Acc, B-Acc, and diversity. This suggests that a simple CoT, lacking an explicit and verifiable reasoning path, is insufficient to effectively elicit the model’s parametric knowledge for simulating different users.

CLAIMSIM predictions are more diverse and comparatively more align to true distributions.

Base Model	Variant	Gender		Politics		Religion	
		Acc	B-Acc	Acc	B-Acc	Acc	B-Acc
GPT-4O-MINI	DIRECT PROMPTING	0.40(± 0.01)	0.66 (± 0.02)	0.41 (± 0.05)	0.73 (± 0.01)	0.28(± 0.04)	0.52(± 0.03)
	CoT	0.40(± 0.01)	0.66 (± 0.02)	0.41 (± 0.04)	0.73 (± 0.01)	0.34(± 0.05)	0.56(± 0.03)
	CLAIMSIM	0.30(± 0.04)	0.66 (± 0.02)	0.41 (± 0.05)	0.72(± 0.02)	0.32(± 0.04)	0.56(± 0.03)
QWEN 3	DIRECT PROMPTING	0.40(± 0.03)	0.66 (± 0.02)	0.37(± 0.02)	0.69(± 0.02)	0.33(± 0.02)	0.59(± 0.03)
	CoT	0.41(± 0.01)	0.66 (± 0.01)	0.37(± 0.03)	0.68(± 0.02)	0.31(± 0.03)	0.59(± 0.02)
	CLAIMSIM	0.36(± 0.04)	0.66 (± 0.02)	0.37(± 0.03)	0.67(± 0.02)	0.30(± 0.01)	0.60(± 0.03)
LLAMA 4	DIRECT PROMPTING	0.33(± 0.04)	0.65(± 0.02)	0.35(± 0.03)	0.69(± 0.01)	0.33(± 0.00)	0.58(± 0.03)
	CoT	0.42 (± 0.02)	0.65(± 0.02)	0.32(± 0.03)	0.70(± 0.01)	0.31(± 0.02)	0.57(± 0.03)
	CLAIMSIM	0.40(± 0.02)	0.61(± 0.03)	0.29(± 0.02)	0.63(± 0.02)	0.36 (± 0.04)	0.62 (± 0.02)

Table 1: Performance comparison between base LLMs in DIRECT PROMPTING mode, with CoT prompting, and with CLAIMSIM built on each LLM. Results are shown as mean \pm standard deviation (in smaller font). The best performance in each domain is highlighted in **bold**.

Base Model	Variant	Gender	Politics	Religion
GPT-4O-MINI	DIRECT PROMPTING	0.56	0.63	0.70
	CoT	0.54	0.59	0.68
	CLAIMSIM	0.47	0.96	0.47
QWEN 3	DIRECT PROMPTING	0.34	0.36	0.42
	CoT	0.28	0.38	0.45
	CLAIMSIM	0.59	0.53	0.40
LLAMA 4	DIRECT PROMPTING	0.83	0.76	0.53
	CoT	0.65	0.79	0.59
	CLAIMSIM	0.62	0.68	0.62

Table 2: Wasserstein distances between each method’s predictions and the ground truth across different base models and domains. Lower scores indicate greater distributional similarity with the ground truth over ordinal labels. The best performance for each backbone model within each domain is highlighted in **bold**.

As shown in Figure 2, the answer distributions generated by CoT are heavily concentrated around specific options, with this effect most pronounced in the LLAMA 4 model. Table 2 further demonstrates that, in most cases, CLAIMSIM achieves a closer alignment with the ground truth distribution. This suggests that DIRECT PROMPTING and CoT tend to reflect a unified viewpoint, even when provided with diverse demographic features. In contrast, CLAIMSIM produces more diverse answer distributions by explicitly instructing LLMs to elicit multiple perspectives through diverse claim generation.

Nuanced difference between CLAIMSIM and CoT within the opinion category. As shown in Table 1, the performance gap between CoT and CLAIMSIM is notably smaller under binary labels (B-Acc) compared to finer-grained labels (Acc). This suggests that prompting LLMs to generate diverse claims primarily results in nuanced shifts within the same opinion category, e.g., from

Strongly Disagree to Disagree. Fundamentally altering the model’s stance may require direct interventions, such as instruction fine-tuning or RLHF.

None of the approaches provide accurate user simulations. All three methods achieve slightly above random accuracy across different backbone LLMs (Table 1), rendering the inherent difficulty of the task. Even though CLAIMSIM enhances diversity and achieves better distribution alignment, these improvements are not consistently stable across different tasks (Table 2). In the next section, we conduct a fine-grained analysis with case studies to better understand two key limitations.

3.4 Why LLMs struggle to simulate users?

LLMs fail to reason over conflicting evidence about opinions. As shown in Figure 3, although CLAIMSIM elicits conflicting claims based on the sex and education features, LLMs fail to reason over these contradictions and generate the expected response for the target user. This remains a challenging research problem, as it requires LLMs to capture subtle relationships between demographic attributes. For example, recognizing that individuals born in the 1970s with a mid-level education may be slightly less opposed to the given view.

LLMs produce unified viewpoints regardless demographic attributes. As shown in Figure 4, LLMs consistently generate claims that agree with the survey question, even when prompted with varying demographic profiles. Our manual analysis reveals that this pattern occurs in 50% of the survey questions tested. While RLHF significantly helps align LLMs with widely accepted moral values, it also fundamentally limits their ability to simulate users whose viewpoints diverge from these norms.

Question: A university education is more important for a boy than for a girl. (A) Strongly Agree (B) Agree (C) Disagree (D) Strongly Disagree
Demographic Information: Sex - Male; Educational level - Middle; Year of birth - 1970s ...
True Label: (B) Agree
CoT Prediction: Answer: (D) Strongly Disagree Reasoning: As a male from 1970s with a middle education, I value education equally for both genders, despite past societal norms favoring boys.
CLAIMSIM Prediction: Answer: (D) Disagree Claim for sex: Male Boys benefit more from a university education due to societal expectations. Claim for education: Middle Equal importance of university education for both boys and girls is supported. Reasoning: As a male with a middle-level education, I believe both boys and girls deserve equal access to university education, regardless of societal expectations.

Figure 3: A case study showing LLMs fail to reason over conflicting evidence about opinions.

Question: Being a housewife is just as fulfilling as working for pay. (A) Strongly Agree (B) Agree (C) Disagree (D) Strongly Disagree
Demographic Information: Sex - Female; Educational level - Lower; Religion: Protestant ...
True Label: (B) Agree
CoT Prediction: Answer: (A) Strongly Agree Reasoning: As a North American Protestant woman with lower educational attainment, I find deep fulfillment in managing my home.
CLAIMSIM Prediction: Answer: (A) Strongly Agree Claim for sex: Female Being a housewife can be fulfilling by highlighting personal flexibility. Claim for education: Lower Housewifery is a legitimate choice, emphasizing its emotional rewards. Reasoning: As a Protestant woman with lower education, being a housewife has shown me the deep fulfillment through nurturing family.

Figure 4: A case study indicating LLMs produce unified viewpoints regardless of attributes.

LLMs fail to generate diverse claims for certain questions due to inherent bias. Across five sampling rounds, LLMs successfully generate diverse claims for 60% of questions in the gender domain, 15% in the politics domain, and all questions in the religion domain. However, inherent biases limit diversity for some questions. For example, when asked, “Do you think that your country’s government should or should not have the right to keep people under video surveillance in public areas?”, the models consistently produce affirmative claims across all demographic groups.

4 Related Work

Large Language Models (LLMs) have shown strong potential in human simulation tasks (Sreedhar and Chilton, 2024; Binz et al., 2025), especially in the context of social science research

involving survey questions that capture personal opinions and group-level perspectives (Cao et al., 2025; Sun et al., 2024; Kim and Lee, 2024). For instance, Cao et al. (2025) explore how LLMs can simulate national-level responses using supervised fine-tuning (SFT) to align model outputs with real-world distributions. At the individual level, Park et al. (2024) propose a generative agent framework that integrates demographic profiles and interview scripts to answer survey questions. In contrast, our work focuses on investigating whether LLMs can elicit internal knowledge purely based on individual demographic attributes.

5 Discussion

During the elicitation of contradicting claims, we observed that the diversity of the claims are still largely constrained to LLMs’ parametric knowledge. Future research shall look into effective ways for diverse claim elicitations, such as prompt optimizations (Pryzant et al., 2023). We also note that answer accuracy is not the only golden metric, we found that a higher overall accuracy does not guarantee a well-aligned answer distributions. In fact, we observed CoT and CLAIMSIM are both insufficient to reflect opinions from nuanced demographic groups, highlighting the needs for further explorations on how to balance between contradicting claims and reasoning accuracy.

Limitations

In this work, we did not study the consistency of individual-level opinions and simulation of decision-making across multiple domains, but we denote this as important perspective to realize user simulation that shall be investigated in future work. We note that our experiments explored three representative and recent LLMs, but did not include reasoning models such as OpenAI O3, Deepseek R1. This involves substantially longer inference time and API cost, but may benefit both methods with more comprehensive claim generation and advanced reasoning capability. Also due to resource constraints, we did not explore fine-tuning or RLHF existing LLMs, despite we hypothesize that simple post-training approaches cannot tackle these limitations. We also notice that scaling up the domains and demographic nuances of simulated users could also bring new findings. We leave these comparisons and experiments for future work.

Ethics Statement

This work investigates the reliability and associated risks of large language models (LLMs) in simulating user opinions. We identify key limitations in existing approaches and demonstrate improvements in opinion diversity using our proposed method, CLAIMSIM. All experiments were conducted using publicly available LLMs or APIs, and no systems were deployed in real-world settings. Given the broader applications and societal implications of this task, we recognize several ethical concerns, including the risks of hallucinated content, overconfident claims, and the amplification of harmful biases. These issues, if left unaddressed, could lead to significant misuse or harm. Our approach, CLAIMSIM, is designed to mitigate these risks by promoting diversity and reducing bias, with the goal of supporting the development of safer and more trustworthy LLM applications in Social NLP. However, we acknowledge that limitations remain, and further investigation is necessary to fully understand and address the ethical and practical challenges posed by this work.

Acknowledgements

This work was supported in part through NYU Shanghai Center for Data Science, STCSM 23YF1430300 and NYU IT High Performance Computing resources, services, and staff expertise.

References

- Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*, pages 337–371. PMLR.
- Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2024. Sensitivity, performance, robustness: Deconstructing the effect of sociodemographic prompting. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2589–2615.
- Marcel Binz, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, Peter Dayan, Can Demircan, Maria K. Eckstein, Noémi Éltető, Thomas L. Griffiths, Susanne Haridi, Akshay K. Jagadish, Li Ji-An, Alexander Kipnis, Sreejan Kumar, Tobias Ludwig, Marvin Mathony, Marcelo Mattar, Alireza Modirshanechi, Surabhi S. Nath, Joshua C. Peterson, Milena Rmus, Evan M. Russek, Tankred Saanum, Johannes A. Schubert, Luca M. Schulze Buschoff, Nishad Singh, Xin Sui, Mirko Thalmann, Fabian Theis, Vuong Truong, Vishaal Udandara, Konstantinos Voudouris, Robert Wilson, Kristin Witte, Shuchen Wu, Dirk Wulff, Huadong Xiong, and Eric Schulz. 2025. *Centaur: a foundation model of human cognition*. Preprint, arXiv:2410.20268.
- Marcel Binz and Eric Schulz. 2024. *Turning large language models into cognitive models*. In *The Twelfth International Conference on Learning Representations*.
- James Brand, Ayelet Israeli, and Donald Ngwe. 2024. *Using llms for market research*. Working Paper 23-062, Harvard Business School Marketing Unit. Last revised: 29 July 2024. 46 pages.
- 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. *arXiv preprint arXiv:2502.07068*.
- Tiancheng Hu, Yara Kyrychenko, Steve Rathje, Nigel Collier, Sander van der Linden, and Jon Roozenbeek. 2024. Generative language models exhibit social identity biases. *arxiv [preprint](2023)*.
- Junsol Kim and Byungkyu Lee. 2024. *Ai-augmented surveys: Leveraging large language models and surveys for opinion prediction*. Preprint, arXiv:2305.09620.
- Lisa Messeri and MJ Crockett. 2024. Artificial intelligence and illusions of understanding in scientific research. *Nature*, 627(8002):49–58.
- Meta AI. 2025. *The llama 4 herd: The beginning of a new era of natively multimodal ai innovation*. Official blog post introducing Llama 4 Scout, Llama 4 Maverick, and Llama 4 Behemoth models.
- OpenAI. 2024. Openai api reference — session object. https://platform.openai.com/docs/api-reference/realtime-sessions/session_object. Accessed: 2025-06-30.
- OpenAI. 2025. *Introducing gpt-4.1 in the api*. <https://openai.com/index/gpt-4-1/>. A new series of GPT models featuring major improvements on coding, instruction following, and long context—plus our first-ever nano model.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS ’22*, Red Hook, NY, USA. Curran Associates Inc.

- Joon Sung Park, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S. Bernstein. 2024. [Generative agent simulations of 1,000 people](#). *arXiv:2411.10109*.
- Reid Pryzant, Dan Iter, Jerry Li, Yin Lee, Chenguang Zhu, and Michael Zeng. 2023. [Automatic prompt optimization with “gradient descent” and beam search](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7957–7968, Singapore. Association for Computational Linguistics.
- Qwen Team. 2025. [Qwen3: Think deeper, act faster](#). Official blog post introducing Qwen3 models and their technical details.
- Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. [Whose opinions do language models reflect?](#) *arXiv preprint arXiv:2303.17548*.
- Patrick Schramowski, Ceyda Turan, Nora Andersen, Xiang Shao, Bernhard Schölkopf, Francis Herbert, Arun Mahendran, Oliver Hinz, and Kristian Kersting. 2022. [Large language models have general value-based biases](#). *Patterns*, 3(7):100537.
- Karthik Sreedhar and Lydia Chilton. 2024. [Simulating human strategic behavior: Comparing single and multi-agent llms](#). *Preprint*, arXiv:2402.08189.
- Seungjong Sun, Eungu Lee, Dongyan Nan, Xiangying Zhao, Wonbyung Lee, Bernard J Jansen, and Jang Hyun Kim. 2024. [Random silicon sampling: Simulating human sub-population opinion using a large language model based on group-level demographic information](#). *arXiv preprint arXiv:2402.18144*.
- Cédric Villani et al. 2008. *Optimal transport: old and new*, volume 338. Springer.
- Angelina Wang, Jamie Morgenstern, and John P Dickerson. 2025. [Large language models that replace human participants can harmfully misportray and flatten identity groups](#). *Nature Machine Intelligence*, pages 1–12.
- Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael Lyu. 2024. [Not all countries celebrate thanksgiving: On the cultural dominance in large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6349–6384.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). *Advances in Neural Information Processing Systems*, 35:24824–24837.
- World Values Survey. 2020. [World values survey: Wave 7 \(2017-2020\)](#). Accessed: 2024-09-11.

Base Model	Sample Count	Gender		Politics		Religion	
		Acc	B-Acc	Acc	B-Acc	Acc	B-Acc
GPT-4O-MINI	100 Users	0.40	0.66	0.41	0.73	0.28	0.52
	200 Users	0.41	0.68	0.38	0.68	0.31	0.52
QWEN 3	100 Users	0.40	0.66	0.37	0.69	0.33	0.59
	200 Users	0.40	0.67	0.35	0.70	0.32	0.59
LLAMA 4	100 Users	0.33	0.65	0.35	0.69	0.33	0.58
	200 Users	0.32	0.63	0.34	0.67	0.36	0.57

Table 3: Acc and B-Acc across three domains (Gender, Politics, Religion) for 100 and 200 sampled users under the DIRECT PROMPTING setting, evaluated on three backbone models.

Base Model	Metric	Gender		Politics		Religion	
		t	p	t	p	t	p
GPT-4O-MINI	Acc	-0.58	0.591	0.66	0.535	-2.05	0.133
	B-Acc	-1.09	0.336	1.89	0.107	-0.06	0.958
QWEN 3	Acc	0.18	0.867	0.58	0.581	0.60	0.592
	B-Acc	-0.23	0.830	-0.61	0.562	-0.14	0.898
LLAMA 4	Acc	0.27	0.802	1.32	0.236	-1.27	0.294
	B-Acc	0.77	0.482	2.28	0.062	0.05	0.962

Table 4: p-test comparison of Exact Accuracy and Binary Accuracy between sampling 100 users and 200 users across models and domains under DIRECT PROMPTING setting.

A Sampling Count Verification

To confirm that a sample size of 100 users does not bias the experimental results, we randomly sample an additional 200 users with diverse demographic information and compare the DIRECT PROMPTING Acc and B-Acc using a p-test (see Table 3 & Table 4).

As shown in Table 4, all Acc and B-Acc comparisons yield p-values greater than 0.05, indicating no significant difference between sampling 100 and 200 users. These results suggest that a sample size of 100 users is sufficient and can be considered a valid choice even when scaling up.

B Distributions Across Three Domains

Figure 5 presents the response distributions of three LLMs across the domains of Gender, Politics, and Religion. Each subplot illustrates how model outputs vary in opinion strength and direction within each domain.

C Analysis on CoT

As mentioned in §3.3, CoT does not improve much compared to DIRECT PROMPTING. In this section we analyze in detail why CoT can not improve much.

Table 6 shows that the reasoning focus of DIRECT PROMPTING and CoT is quite similar, even

though CoT organizes the reasoning into several sub-points. While extensive research has demonstrated that CoT is particularly effective for complex tasks requiring multi-step reasoning, such as solving math problems (Wei et al., 2022), our task (1) does not involve clear hierarchical reasoning steps. As a result, the outputs of DIRECT PROMPTING and CoT are largely comparable; (2) the tested LLMs struggle to autonomously construct effective reasoning processes for user simulation tasks. For example, although LLMs can break the task into steps like *understanding the statement* and *personal context*, they fail to fully integrate all demographic information—such as *sex*—into their reasoning or determine the relative importance of each feature.

D Analysis on CLAIMSIM

Table 7 shows claims generated from individuals with different demographic features through five rounds of sampling. Table 5 reports the percentage of pairwise demographic comparisons that exhibit disparate attitude distributions (Wasserstein distance ≥ 0.05) across three language models (GPT, LLaMA, and QWEN) and five demographic categories (country of birth, educational level, religious denomination, sex, and year of birth) for the gender, government, and religion domains. Notably, all demographic features exhibit disparities exceeding 60%, indicating that LLMs can reflect differences associated with diverse demographic backgrounds to an extent.

E Prompt Template for DIRECT PROMPTING

Below we include the [prompt template](#) for the DIRECT PROMPTING method and an [example prompt](#).

F Prompt Template for CoT

Below we include the [prompt template](#) for the CoT method.

G Prompt Template for CLAIMSIM

Below we include the [prompt template](#) for the CLAIMSIM method, including the [prompt template](#) to generate claims for each of demographic features, generating summaries based on the claims and [prompt template](#) for generating the final responses.

Model	Domain	Country of Birth	Highest Educational	Religious Denomination	Sex	Year of Birth	Row Avg
GPT-4o-MINI	Gender	72.0%	66.7%	58.1%	60.0%	63.8%	64.1%
	Government	61.9%	57.1%	48.3%	42.9%	55.1%	53.1%
	Religion	56.7%	58.3%	57.1%	25.0%	60.7%	51.6%
LLAMA 4	Gender	66.7%	73.3%	79.0%	60.0%	53.3%	66.5%
	Government	74.0%	61.9%	76.9%	42.9%	51.7%	61.5%
	Religion	75.0%	75.0%	72.6%	75.0%	42.9%	68.1%
QWEN 3	Gender	82.2%	93.3%	92.4%	60.0%	84.8%	82.5%
	Government	74.6%	81.0%	78.2%	85.7%	70.1%	77.9%
	Religion	86.1%	58.3%	73.8%	50.0%	71.4%	67.9%
Col Avg		72.1%	69.4%	70.7%	55.7%	61.5%	65.9%

Table 5: **Wasserstein Disparity Matrix across Models and Domains.** This table summarizes the degree of demographic disparity identified across different language models (GPT, LLAMA, and QWEN) and domains (Gender, Government, and Religion). Each cell reports the percentage of pairwise demographic comparisons within a given category that are considered disparate, defined as having a Wasserstein distance ≥ 0.05 . Higher percentages indicate greater divergence in response distributions between demographic groups.

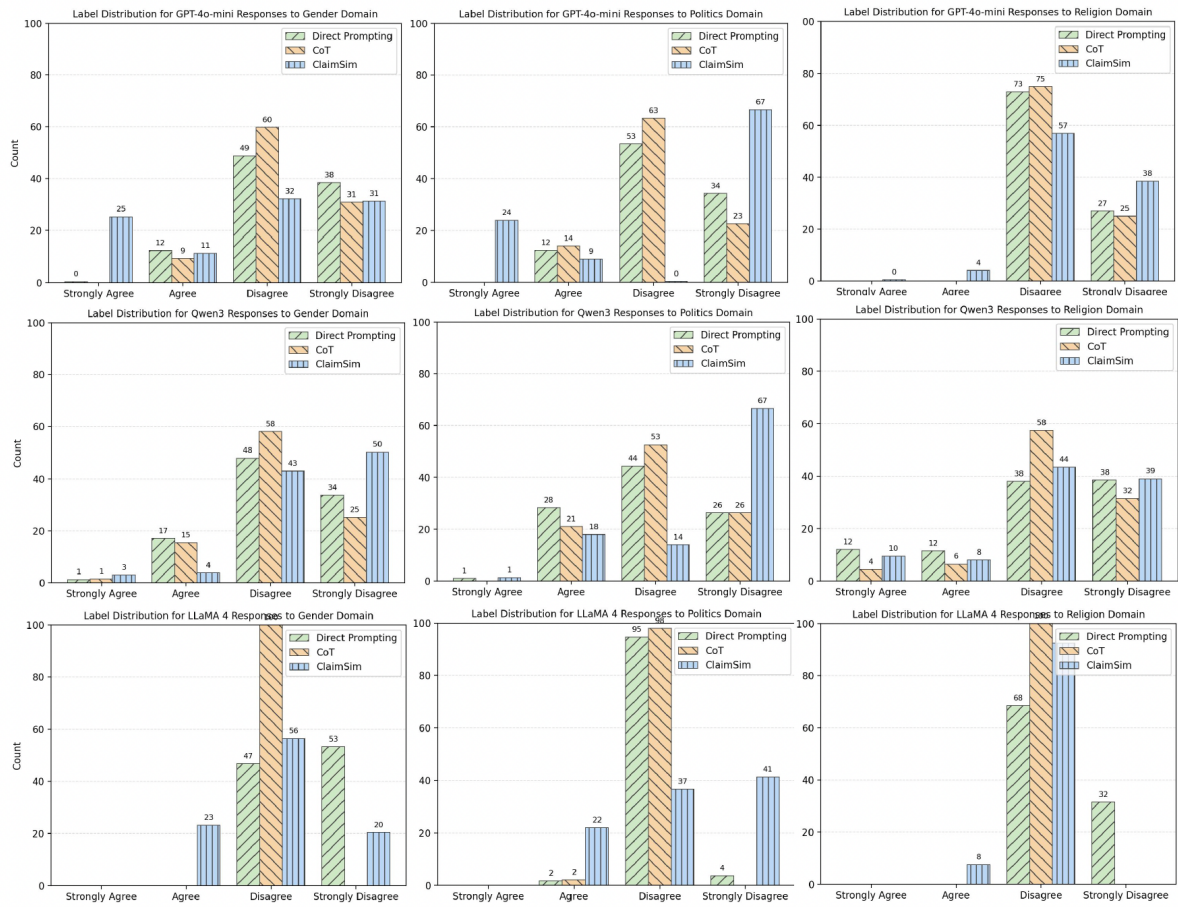


Figure 5: Response distributions of three LLMs across the Gender, Politics, and Religion domains.

Category	Content
Demographic Information	Sex: Male; Year of birth: 1950s; Age: 65 and over; Country of birth: Western Europe; Marital status: Married; Highest educational level: Higher; Occupational group: Professional/Managerial; Religious denomination: No religion
Question	When a mother works for pay, the children suffer. Do you strongly agree, agree, disagree, or strongly disagree?
Response (DIRECT PROMPTING)	<p>Label: Strongly disagree</p> <p>Reasoning: As a retired professional with higher education, I value both parents' contributions. A mother's work brings financial stability and models independence and work ethic. The belief that children suffer is outdated; dual-income families can offer children diverse, positive experiences.</p>
Response (CoT)	<p>Label: Disagree</p> <p>Reasoning: 1. <i>Understanding the statement:</i> Suggests maternal employment harms children—an outdated view. 2. <i>Personal context:</i> I grew up with working parents and value hard work and education. 3. <i>Cultural shifts:</i> Gender roles evolved; children can thrive in dual-income homes. 4. <i>Benefits of maternal employment:</i> Provides stability, models work ethic, and teaches life skills. 5. <i>Personal experience:</i> I've seen friends balance work and family successfully. 6. <i>Conclusion:</i> Maternal work does not inherently harm children; I disagree with the statement.</p>
Response (CLAIMSIM)	<p>Claims Summary: 1. <i>Male:</i> The respondent expresses both positive and negative views—four claims highlight benefits of maternal employment, while one suggests children suffer due to reduced care. 2. <i>Born in 1950s:</i> All five claims emphasize negative impacts, focusing on reduced parental attention and emotional connection. 3. <i>Higher educational level:</i> All five claims support the perspective that maternal work has beneficial impacts on children, with no claims opposing this view. 4. <i>No religion:</i> Five claims support this perspective, highlighting that working mothers can provide better opportunities and that the belief that children suffer from maternal employment is overly simplistic. 5. <i>Born in Western Europe:</i> The respondent presents a <u>balanced view</u> on maternal employment, recognizing both concerns about reduced emotional support and quality time (2 claims) and the benefits of greater financial stability and opportunities for children (3 claims).</p> <p>Label: Disagree</p> <p>Reasoning: Drawing on my demographic background, I balanced the tension in the claims and weighed the benefits of maternal employment more heavily. 1. <i>Highest education level and sex:</i> My education encourages a nuanced view; I see both sides but value the benefits of working mothers, like financial stability and role modeling. 2. <i>Racial/ethnic group and country of birth:</i> Western European context supports gender equality, aligning with claims that maternal employment helps child development. 3. <i>Integration of claims:</i> Although reduced maternal attention is a concern, overall the positive impacts outweigh drawbacks, leading me to disagree with the statement.</p>

Table 6: Example responses under DIRECT PROMPTING, CoT, and CLAIMSIM modes (this example includes only a subset of personal demographic information due to space constraints). Detailed claims are shown in Table 7

Demographic Feature	Claims
Male	<ol style="list-style-type: none"> 1. When a mother works for pay, the children do not necessarily suffer; in fact, they can benefit from her employment. 2. Children benefit from having a working mother. 3. Children can thrive even when their mothers work for pay. 4. The assertion that when a mother works for pay, the children suffer is an oversimplification that does not account for the positive impacts of maternal employment on family dynamics and child development. 5. When a mother works for pay, the children suffer due to the lack of direct maternal care and attention.
Born in 1950s	<ol style="list-style-type: none"> 1. When a mother works for pay, the children suffer due to a lack of parental attention and support. 2. When a mother works for pay, the children suffer in terms of emotional connection and time spent together. 3. Working mothers can negatively impact their children's emotional and social development. 4. When a mother works for pay, the children suffer due to reduced time and attention at home. 5. When a mother works for pay, the children suffer due to reduced maternal presence and support.
Higher Educational Level	<ol style="list-style-type: none"> 1. The assertion that when a mother works for pay, the children suffer is an oversimplification that ignores the positive impacts of maternal employment on both the mother and her children. 2. When a mother works for pay, the children can benefit from increased financial stability and role modeling of work ethics. 3. A mother's paid work does not inherently harm children; in fact, it can provide them with better opportunities. 4. Working mothers can provide better opportunities for their children. 5. A mother's work for pay can positively influence her children's development and wellbeing.
No Religion	<ol style="list-style-type: none"> 1. A mother's decision to work for pay does not necessarily mean that her children suffer; in fact, it can have positive effects on their development and wellbeing. 2. Working mothers can provide better opportunities for their children. 3. The belief that children suffer when their mothers work for pay is overly simplistic and does not account for the diverse realities of modern families. 4. Working mothers can positively influence their children's development and wellbeing. 5. Children can thrive even when their mothers work for pay.
Born in Western Europe	<ol style="list-style-type: none"> 1. The claim that when a mother works for pay, the children suffer is overly simplistic and does not consider the multifaceted benefits of maternal employment. 2. Mothers working for pay can provide better financial stability and opportunities for their children. 3. Working mothers can provide better opportunities for their children. 4. When a mother works for pay, the children may suffer in terms of emotional support and quality time spent together. 5. When a mother works for pay, it can lead to negative consequences for children, including reduced quality time and emotional support.

Table 7: Example claims for each of five demographic features in response of the survey question "Do you agree that when a mother works for pay, the children suffer."

Prompt Template for DIRECT PROMPTING

Personal Demographic Information

The respondent's demographic details are as follows:

```
{% for category, value in demo_infos.items() %}  
- {{ category }}: {{ value }}  
{% endfor %}
```

Task

Imagine you are the respondent. Based on your demographic background, answer the following question under the topic of {{ domain }}.

Instruction

```
{{ instruction }}
```

Question

```
{{ question }}
```

Label Choices

You must choose **exactly one** label from the options below:

```
{{ labels }}
```

Response Format

-Label: The selected label from the provided choices.

-Reasoning: An explanation leading to your choice.

Prompt Example for DIRECT PROMPTING

Personal Demographic Information

The respondent's demographic details are as follows:

- **Sex:** Male
- **Year of birth:** 1940s
- **Age:** 65 and over
- **Respondent immigrant:** I am an immigrant to this country (born outside this country)
- **Mother immigrant:** Immigrant
- **Father immigrant:** Immigrant
- **Country of birth: Respondent:** East Asia
- **Respondent citizen:** yes
- **Number of people in household:** 1
- **Do you live with your parents:** Not live with parents
- **Marital status:** Married
- **How many children do you have:** 1 child
- **Highest educational level: Respondent:** Higher
- **Employment status:** Retired/pensioned
- **Employment status - Respondent's Spouse:** Full time (30 hours a week or more)
- **Respondent - Occupational group:** Professional/Managerial
- **Sector of employment:** Private business or industry
- **Are you the chief wage earner in your house:** Yes
- **Family savings during past year:** Save money
- **Social class:** Lower class
- **Income level:** Low
- **Religious denominations - major groups:** Protestant

Task

Imagine you are the respondent. Based on your demographic background, answer the following question under the topic of **gender**.

Instruction

For each of the following statements I read out, can you tell me how strongly you agree or disagree with each. Do you strongly agree, agree, disagree, or strongly disagree?

Question

When a mother works for pay, the children suffer.

Label Choices

You must choose **exactly one** label from the options below:

['strongly agree', 'agree', 'disagree', 'strongly disagree']

Response Format

- Label: The selected label from the provided choices.
- Reasoning: An explanation leading to your choice.

Prompt Template for CoT

Personal Demographic Information

The respondent's demographic details are as follows:

```
{% for category, value in demo_infos.items() %}  
- {{ category }}: {{ value }}  
{% endfor %}
```

Task

Imagine you are the respondent. Based on your demographic background, thoughtfully answer the following question under the topic of `{{ domain }}`.

Before selecting your final answer, you must **reason step-by-step** to demonstrate your thought process.

Instruction

```
{{ instruction }}
```

Question

```
{{ question }}
```

Label Choices

You must choose **exactly one** label from the options below:

```
{{ labels }}
```

Response Format

- Label: The selected label from the provided choices.
- Reasoning: A detailed step-by-step explanation leading to your choice, following a chain-of-thought approach.

Important: If you do not provide a detailed reasoning step-by-step, your answer will be considered incomplete.

Claim Generation Prompt

Instruction

Suppose you are a respondent with the following demographic information:

- {{ feature_category }}: {{ feature_label }}

Related Question

- Question: {{ question }}

Task

Please provide all the viewpoints you can think of under the related question. Remember to ground every single opinion in **your** specific background and lived experience. Take reality and ideals into account. For each opinion, include:

- Topic (e.g., the keyword of the opinion)
- Claim (A clear expression of the viewpoint)
- Evidence_for_claim (e.g., the evidence to support the claim)
- Counterpoint (the opposite of the viewpoint)
- Evidence_for_counterpoint (e.g., the evidence to support the counterpoint)

List **1** most representative opinion in plain text.

Output Format (Required)

- **Claim:** Your claim for this question.

Prompt for Summary Generation

Task

Below are some claims from the respondent with the following demographic feature:

- {{ feature_category }}: {{ feature_label }}

Please provide a concise summary that captures the key perspectives expressed in the claims.

Claims

{% for claim in claims %}

- **Claim:** {{ claim }}

{% endfor %}

Output Format (Required)

- Summary: Provide a 2-3 sentence synthesis of the respondent's views, clearly identifying key themes, contradictions, or tensions. Explicitly state how many claims support one perspective versus how many support an opposing or contrasting view, if applicable.

Prompt Template for CLAIMSIM

Demographic Profile

Below is a simulated demographic profile. Please respond as if you belong to this background:

```
{% for category, value in demo_infos.items() %}  
- {{ category }}: {{ value }}  
{% endfor %}
```

Context

You will evaluate claims and counterpoints that reflect opinions or beliefs people might hold in the field of `{{ domain }}`. Each pair is designed to capture a possible tension or debate that may arise based on demographic perspectives.

These statements are *hypothetical and intentionally diverse* to explore how views might vary across backgrounds. Your task is not to judge them by factual accuracy, but to engage thoughtfully based on your assigned profile.

Statements

For each of the following, consider both the claim and the counterpoint:

```
{% for claim, counterpoint in claims.items() %}  
- Claim: {{ claim }}  
- Counterpoint: {{ counterpoint }}  
{% endfor %}
```

Task

You are asked to select the position that would most closely align with the simulated demographic perspective above. This is a reasoned choice based on how someone from this profile might respond. In doing so, carefully consider how the specific claims made in the prompt may influence their reasoning. At the same time, critically reflect on potential counterpoints—how someone from this demographic might still be persuaded by alternative views. Your answer should weigh these tensions and offer a thoughtful justification.

Instruction

```
{{ instruction }}
```

Question

```
{{ question }}
```

Label Choices

Choose **exactly one** of the following:

```
{{ labels }}
```

Response Format (Required)

- **Label:** your selected label from above
- **Reasoning:** Step-by-step explanation of how the claims, counterpoints and simulated demographic background influence the choice. Be specific and avoid generic justifications.

> **Warning:** Incomplete responses without detailed reasoning will be considered invalid for this task.