SocioBench: Modeling Human Behavior in Sociological Surveys with Large Language Models

Jia Wang[12], **Ziyu Zhao**[13], **Tingjuntao Ni**[14], **Zhongyu Wei**[15]*
[1]Shanghai Innovation Institute, [2]Tongji University, [3]Zhejiang University,
[4]Shanghai Jiao Tong University, [5]Fudan University
jiawang@tongji.edu.cn, zywei@fudan.edu.cn

Abstract

Large language models (LLMs) show strong potential for simulating human social behaviors and interactions, yet lack large-scale, systematically constructed benchmarks for evaluating their alignment with real-world social attitudes. To bridge this gap, we introduce SocioBench—a comprehensive benchmark derived from the annually collected, standardized survey data of the *International Social Survey Programme (ISSP)*. The benchmark aggregates over 480,000 real respondent records from more than 30 countries, spanning 10 sociological domains and over 40 demographic attributes. Our experiments indicate that LLMs achieve only 30–40% accuracy when simulating individuals in complex survey scenarios, with statistically significant differences across domains and demographic subgroups. These findings highlight several limitations of current LLMs in survey scenarios, including insufficient individual-level data coverage, inadequate scenario diversity, and missing group-level modeling. We have open-sourced **SocioBench** at https://github.com/JiaWANG-TJ/SocioBench.

1 Introduction

As the LLMs advance in generating natural language (Min et al., 2023; Karanikolas et al., 2024; Gao et al., 2025), simulating cognitive processes (Niu et al., 2024; Subramonyam et al., 2024; Ren et al., 2025; Azaria et al., 2023; Chen, 2024), and engaging in complex dialogues (Mou et al., 2024b; Li et al., 2024), their potential applications in the social sciences are becoming increasingly evident (Anthis et al., 2025; Aher et al., 2023; Chen et al., 2024). Beyond analyzing large-scale textual data, LLMs can function as "computational agents" that simulate human behavior (Liu et al., 2024; Wang et al., 2025) and decision-making (Sun et al., 2025; Li et al., 2025), enabling social experiments and surveys (Zhang et al., 2025; Leng and Yuan, 2023; Mou et al., 2024a) that are difficult to conduct in real-world settings due to ethical, logistical, or financial constraints (Park et al., 2023). Existing research has primarily focused on micro-level social capabilities such as persona consistency, linguistic style, and personality traits, or on group-level tasks like social reasoning, social bias identification, and multi-agent cooperation (Ji et al., 2025; Strachan et al., 2024; Li et al., 2023). Although benchmarks such as OpinionQA (Santurkar et al., 2023) have made important strides in evaluating these aspects, few have systematically assessed LLMs' ability to reflect macro-level social attitudes and cross-cultural differences.

To bridge this gap, we develop **SocioBench**, a large-scale, cross-national benchmark for simulating human behavior in social survey scenarios. The benchmark is built upon the ISSP's (Group, 2015, 2016b,a, 2017, 2018, 2019, 2020, 2022, 2023, 2024) standardized questionnaires and 481,629 authentic respondent records, and it covers 10 research domains: *Citizenship, Environment, Family and Changing Gender Roles, Health and Health Care, National Identity, Religion, Role of Government, Social Inequality, Social Networks, and Work Orientations*. Figure 1 shows an overview of the pipeline for constructing SocioBench.

2 SocioBench Curation

Dataset Statistics. SocioBench is built upon the ISSP, a long-standing, international collaborative project that annually collects standardized data on social attitudes, with its data archive maintained by the GESIS – Leibniz Institute for the Social Sciences[1]. SocioBench covers 10 sociological domains across more than 30 countries. The full version, SocioBench-Full, comprises 481,629 respondents, with each respondent profiled by over

*Corresponding authors.

[1]https://www.gesis.org/en/home

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 26257–26289
November 4-9, 2025

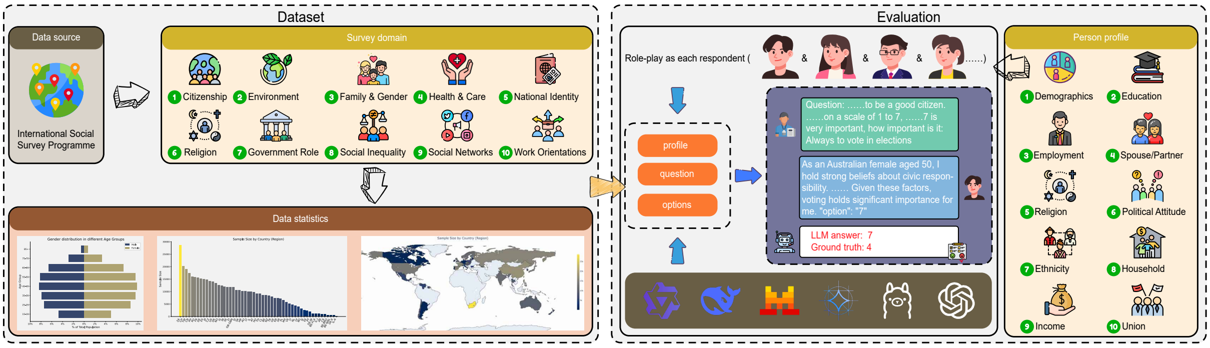


Figure 1: Overview of SocioBench. We first constructed the questionnaire question-answering dataset covering the ten sociological domains of the ISSP, along with the dataset containing ground-truth demographic labels and respondent answers. We then instructed the LLM to answer the survey conditioned on the demographic labels, and evaluated model performance by computing the accuracy between the LLM's responses and the ground-truth answers.

40 demographic features—including age, gender, education level, occupation, income, religious affiliation, and political orientation et al. To enhance computational efficiency, we release sampled versions: **SocioBench-5000**, where the suffix indicates the total number of respondents. Unless otherwise specified, all experiments—excluding those in Section 4 (Data Sampling Ratios Comparison)—are conducted using SocioBench-5000. By default, "SocioBench" refers to this version. The statistical overview is presented in Table 2, while detailed distributions of Q&A and demographic information are available in Appendix A.1 & A.2.

We compare SocioBench with some representative datasets for the analysis of social attitudes and show the results in Table 1. Previous resources adopt partial perspectives, restricted to specific countries, a narrow set of topics, or without demographic diversity. SocioBench, on the contrary, provides a unified benchmark that simultaneously spans languages, domains, demographics, and regions, aligning more closely with real-world social contexts.

Dataset Curation. The SocioBench dataset comprises the questionnaire, respondents' demographic attributes and their responses. The data processing pipeline comprises three steps: first, we filter out open-ended questions and invalid responses (e.g., "Not applicable") in the questionnaire to retain quantifiable closed-ended items. Then, we sample 1% of the data to form **SocioBench-5000** for experiments using a two-stage scheme—stratified by country and then random sampling within each country—in order to balance resource constraints against survey coverage. Examples from SocioBench dataset are provided in Appendix A.3.

3 Experiment Setup

Evaluation Pipeline. The evaluation pipeline engages LLMs in role-playing. A prompt template is designed to mimic authentic survey participation: LLMs are explicitly instructed to adopt the identity of the respondent through embedded demographic profiles (e.g., "You are a 31-year-old Australian woman with a high school to high school education completed, who has a partner, no religious affiliation, and is of Australian ethnicity", see Appendix C). The models then generate answer options accodrding to the sociocultural context.

Comparison Models. We compare state-of-the-art LLMs on SocioBench, including the GPT series, Llama series, Qwen series, Mistral series, and so on (OpenAI et al., 2024; Qwen et al., 2024; Grattafiori et al., 2024; GLM et al., 2024; DeepSeek-AI et al., 2025; Team et al., 2025; Jiang et al., 2024)[2][3].

Evaluation Metrics. To evaluate the alignment of LLMs with real-world social attitudes in SocioBench, we employ the metrics: **Accuracy**. **Accuracy** measures the proportion of model predictions that exactly match the ground-truth responses:

$$\text{Accuracy} = \frac{\sum_{i=1}^{n} \mathbb{I}(y_i^{\text{true}} = y_i^{\text{pred}})}{n} \times 100\% \quad (1)$$

[2] https://github.com/QwenLM/Qwen3
[3] https://github.com/InternLM/InternLM

Dataset	Multilingual?	Multi-domain?	Demographic variables	Multi-regions?
SocioBench (Ours)	✓	✓	✓	✓
SocialBench (Chen et al., 2024)	×	×	×	×
OpinionQA (Santurkar et al., 2023)	×	✓	✓	×
GlobalOpinionQA (Durmus et al., 2023)	✓	✓	×	✓

Table 1: Comparison of social and opinion survey datasets.

Domain	Year	Ctry.	Feat.	Resp.	Q.	Total
Citizen	2014	33	44	500	59	29 500
Enviro	2020	28	45	500	50	25 000
Family	2012	39	45	500	54	27 000
Health	2021	28	45	500	51	25 500
Nat.Ident	2013	35	46	500	60	30 000
Religion	2018	30	46	500	59	29 500
R.Gov	2016	30	46	500	60	30 000
S.Ineq	2019	25	44	500	46	23 000
S.Net	2017	28	47	500	59	29 500
Work	2015	35	47	500	57	28 500
Total	—	—	**408**	**5 000**	**555**	**277 500**

Table 2: Respondent profile information and questionnaire statistics in SocioBench. Abbreviations: Ctry. = Number of countries; Feat. = Number of features; Resp. = Number of respondents; Q. = Number of questions; Tot. = Total. Citizen = Citizenship; Enviro = Environment; Family = Family and Changing Gender Roles; Health = Health and Healthcare; Nat.Ident = National Identity; Religion = Religion; R.Gov = Role of Government; S.Ineq = Social Inequality; S.Net = Social Networks; Work = Work Orientations.

where y_i^{true} and y_i^{pred} denote the true and predicted responses for the i-th sample respectively, n is the total number of valid samples, and $\mathbb{I}(\cdot)$ is an indicator function that equals 1 when the condition is satisfied and 0 otherwise.

Implementation Details. The experiment leverages the vLLM framework to efficiently serve LLMs on 4 NVIDIA H100 GPUs supporting context lengths up to 10,240 tokens. Generation parameters are consistently maintained with a *Temperature* of 0.5, *Top P* of 0.95, *Repetition Penalty* of 1.1.

4 Experimental Results

We conducted extensive experiments, systematically investigating the influence of various factors, including model parameter scale, model family, survey domain, dataset size, and survey rounds in different years. Furthermore, we examine how two factors—whether to enable reasoning and whether to output reasons—affect LLMs' behavioral simulation, and we conduct subgroup analyses based on different demographic information to further explore the bias of the LLM.

The core analyses and findings are presented in this section, while additional results are detailed in Appendix G.

Overall Experimental Results. Our experiments yielded four primary findings. First, when simulating individual behavior in complex social survey scenarios, the accuracy of LLMs is generally 30–40% (see Table 3). This shows the limitations of LLMs in modeling individual behavior.

Second, model performance improves with increasing parameter scale. For instance, within the Qwen2.5 family, Qwen2.5-7B-Instruct, Qwen2.5-32B-Instruct, and Qwen2.5-72B-Instruct achieve average accuracies of 33.35%, 36.03%, and 37.24%, respectively.

Furthermore, across different model families, we find that GLM-4-9B-chat, Qwen2.5-32B-Instruct, and DeepSeek-R1-Distill-Llama-70B emerge as the top-performing models in the < 10B, ~ 30B, and ~ 70B parameter ranges, achieving average accuracies of 35.60%, 36.03%, and 38.52%, respectively.

Finally, model performance varies significantly across different domains. For instance, accuracy peaks at 44.30% in *Citizenship* but is only 36.16% in *Health and Healthcare*. The consistent trend observed across different models is likely due to the uneven data distribution of LLM pre-training corpora. Data scarcity in certain domains results in disparities in the models' semantic comprehension capabilities when addressing sociological issues.

Subgroup Analyses. To analyze biases that may arise when LLMs role-play respondents from different demographic backgrounds, we conducted subgroup analyses using representative models (the Qwen family, the Llama family, and the GPT family). We consider subgroups defined by geographic region (continent), sex, and age range. Moreover, we perform statistical tests to determine whether these labels significantly affect group-level accuracy in behavioral simulation. The detailed data are available in Appendix I.

Model	Citizen	Enviro	Family	Health	Nat.Ident	Religion	R.Gov	S.Ineq	S.Net	Work	Avg.
	Accuracy % (↑)										
BASELINES											
Random Guess	25.93	23.22	21.58	21.24	23.02	20.84	23.64	20.25	18.65	22.99	22.14
GPT-4o	**44.30**	**37.07**	**39.14**	35.33	36.35	40.76	**39.86**	**36.62**	36.69	38.94	38.51
InternLM3-8b-instruct	41.65	33.66	31.05	32.35	34.60	36.61	36.09	32.21	33.96	36.19	34.84
GLM-4-9b-chat	41.81	33.95	31.96	34.13	36.53	37.32	36.03	34.35	31.86	38.10	35.60
Gemma-3-27b-it	40.92	34.63	34.87	30.49	33.84	38.08	35.97	32.60	35.63	38.10	35.51
DeepSeek-R1-Distill-Llama-70B	44.19	35.98	38.11	36.14	37.42	40.65	39.32	35.97	37.38	**39.99**	**38.52**
Mistral-7B-Instruct-v0.3	39.64	32.62	28.16	30.68	32.86	35.85	34.58	30.21	33.81	35.49	33.39
Mixtral-8x22B-Instruct-v0.1	43.10	34.20	34.40	32.38	33.29	37.86	35.89	33.70	37.35	35.11	35.73
Llama-3.1-8B-Instruct	40.43	32.11	31.89	32.21	33.37	36.99	35.27	31.47	34.99	33.39	34.21
Llama-3.3-70B-Instruct	44.03	35.97	38.62	**36.16**	**38.19**	**41.26**	39.19	35.73	36.14	38.80	38.41
Qwen2.5-7B-Instruct	40.90	29.84	30.10	31.82	33.67	36.54	34.80	30.37	33.34	32.18	33.35
Qwen2.5-32B-Instruct	42.54	35.26	34.94	33.20	35.09	37.88	36.32	34.00	34.48	36.57	36.03
Qwen2.5-72B-Instruct	43.59	35.51	36.27	35.90	34.13	39.80	36.56	35.17	**38.06**	37.38	37.24
Qwen3-8B	40.28	32.70	33.07	33.98	33.12	37.58	34.65	30.83	34.38	34.20	34.48
Qwen3-32B	43.60	34.12	34.53	33.53	32.64	38.90	35.52	33.16	35.31	35.25	35.66

Table 3: Comparison of different LLMs across SocioBench. We report the best LLM performance in bold and the second best underlined.

Cross-Continental Analysis: We specifically selected the domains of *Religion* and *Social Inequality* for analysis, see Figure 2. Analysis of variance reveals highly significant differences across continents for all evaluated models (all $p < .001$). Specifically, models exhibit generally lower accuracy when simulating the personas of African respondents compared to those from Europe, North America, and Oceania.

Cross-Gender Analysis: Our analysis of the *Citizenship* and *Family and Changing Gender Roles* domains reveals that the accuracy in simulating female personas is consistently higher than that for male personas. For instance, the respective accuracies are 43.04% ± 1.72% (mean ± standard deviation) and 41.87% ± 1.97% in the *Citizenship*. These findings suggest that training corpora imbalances may lead to female roles being associated with clearer semantic patterns in certain domains, see Figure 14.

Cross-Age Analysis: Our analysis shows that in the *Role of Government* and *Social Networks* domains, the accuracies for the 56–65 and 66-and-over age groups (37.52% ± 2.27% and 37.91% ± 1.45%, respectively) outperform young people, such as the 18–25 and 36–45 age groups. This suggests that these domains are more strongly associated with middle-aged and older populations, or that the social networks and political participation of these groups are more established, thereby enabling LLMs to simulate these demographic groups with greater accuracy, see Figure 15.

Option Distribution in LLMs' Responses. We further conducted a comparative analysis of the distribution of options selected by human respondents and LLMs. The results reveal that although the ground truth exhibits skewed distributions (i.e., options are concentrated in several categories), the LLM-generated responses make this skewness more pronounced, and Llama-3.3-70B-Instruct shows the most marked concentration. Conversely, we observe that Qwen3-32B tends to produce more uniform option distributions. See Appendix F for details.

How do Thinking Modes Shape LLMs' Behavioral Simulation? To analyze how the thinking/reasoning processes affect behavioral simulation in social survey scenarios, we compared Qwen3-8B and Qwen3-32B with and without the thinking mode. The results show that the thinking mode has only a minor effect, yielding slight gains in behavioral simulation accuracy, see Table 11 in the Appendix G. Specifically, the 8B model shows an average improvement of 0.51 percentage point (pp), while the 32B model improves by 0.89 pp. An output example can be found in Appendix D.

Data Sampling Ratios Comparison. To evaluate robustness across different data scales, we further constructed two sub-datasets, SocioBench-10000 and SocioBench-20000, by sampling 2% and 4% of the complete dataset. On SocioBench-5000, SocioBench-10000, and SocioBench-20000, the Llama-3.1-8B-Instruct model achieved aver-

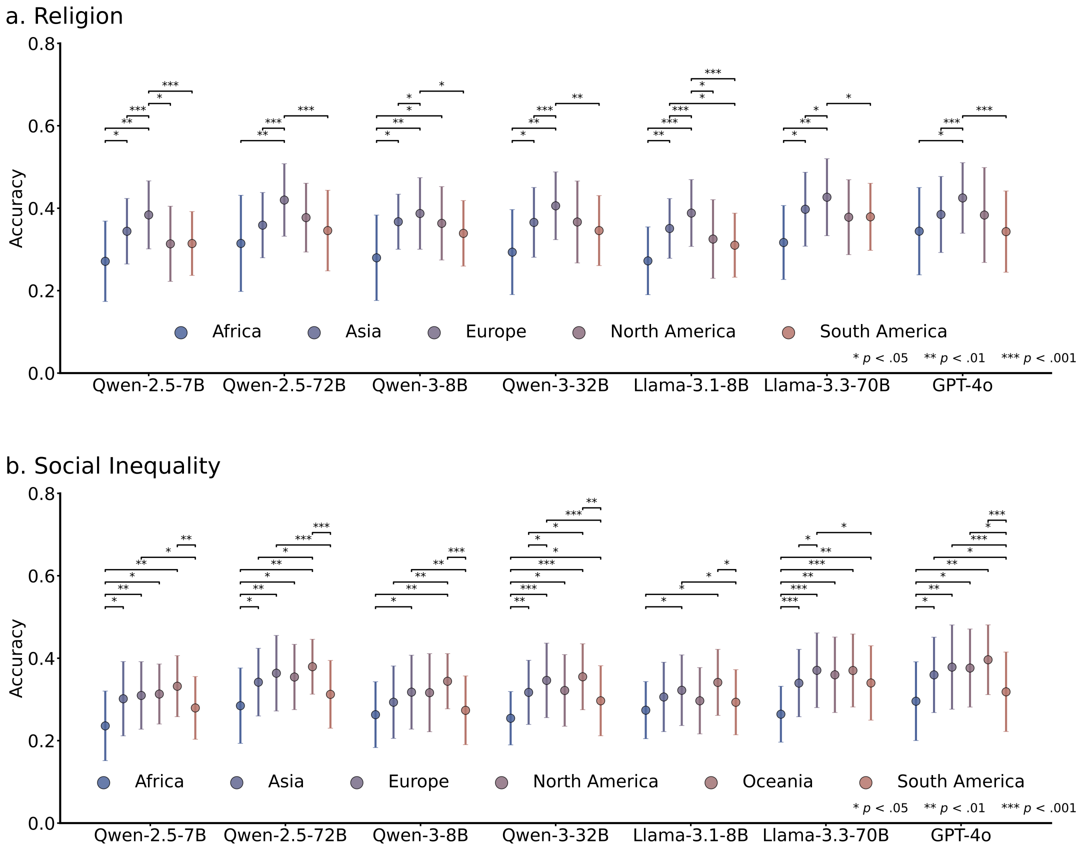


Figure 2: Experimental Results and Significance Analysis of Representative LLMs in the Cross-Continental Subgroup.

age accuracies of 34.21%, 34.28%, and 34.32%, respectively, with a maximum deviation of less than 0.11 pp (see Table 9). These results suggest that small sample sizes yield relatively stable and reliable results.

5 Conclusion

We introduce SocioBench, a cross-cultural benchmark using large-scale real-world sociological survey data to evaluate LLMs' ability to model human behavioral patterns. Through demographic role-play prompts, models generate answers that enable a systematic assessment of alignment with empirically observed social attitudes.

Limitations

Long-Term Data Sustainability. SocioBench relies on the static data of ISSP question–answer pairs and respondent answers. Although these data represent the currently newest survey round results, they cannot track longer-term attitudinal drift.

Evaluation of Dynamism and Openness. The current evaluation relies solely on accuracy, focusing on matching answers at the individual level; and its evaluation of dynamism is insufficient.

Ethic Statement

The SocioBench dataset is based on ISSP[4]. And we contacted the official data provider GESIS (Leibniz Institute for the Social Sciences; isspservice@gesis.org) via email and obtained explicit written permission authorizing the use of the dataset for this study and for publication. Use of the SocioBench must strictly adhere to the data usage requirements of the ISSP and GESIS[5].

Acknowledgements

The research is supported by National Key R&D Program of China (Grant No. 2023YFF1204800) and National Natural Science Foundation of China (Grant No. 62176058). And we sincerely thank the experts at GESIS for their crucial data support and guidance throughout this study.

[4]https://www.issp.org

[5]https://www.gesis.org/en/institute/data-usage-terms

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.

Jacy Reese Anthis, Ryan Liu, Sean M Richardson, Austin C Kozlowski, Bernard Koch, James Evans, Erik Brynjolfsson, and Michael Bernstein. 2025. Llm social simulations are a promising research method. *arXiv preprint arXiv:2504.02234*.

Amos Azaria, Rina Azoulay, and Shulamit Reches. 2023. Chatgpt is a remarkable tool—for experts. *Data Intelligence*, 6:1–49.

Huajun Chen. 2024. Large knowledge model: Perspectives and challenges. *Data Intelligence*, 6:587–620.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, et al. 2024. From persona to personalization: A survey on role-playing language agents. *arXiv preprint arXiv:2404.18231*.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Esin Durmus, Karina Nguyen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

Mingqi Gao, Xinyu Hu, Xunjian Yin, Jie Ruan, Xiao Pu, and Xiaojun Wan. 2025. Llm-based nlg evaluation: Current status and challenges. *Computational Linguistics*, pages 1–28.

Team GLM, :, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Jingyu Sun, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith,

Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mi-

tra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models.

ISSP Research Group. 2015. International social survey programme: National identity iii - issp 2013. GESIS Data Archive, Cologne. ZA5950 Data file Version 2.0.0, https://doi.org/10.4232/1.12312.

ISSP Research Group. 2016a. International social survey programme: Citizenship ii - issp 2014. GESIS Data Archive, Cologne. ZA6670 Data file Version 2.0.0, https://doi.org/10.4232/1.12590.

ISSP Research Group. 2016b. International social survey programme: Family and changing gender roles iv - issp 2012. GESIS Data Archive, Cologne. ZA5900 Data file Version 4.0.0, https://doi.org/10.4232/1.12661.

ISSP Research Group. 2017. International social survey programme: Work orientations iv - issp 2015. GESIS Data Archive, Cologne. ZA6770 Data file Version 2.1.0, https://doi.org/10.4232/1.12848.

ISSP Research Group. 2018. International social survey programme: Role of government v - issp 2016. GESIS Data Archive, Cologne. ZA6900 Data file Version 2.0.0, https://doi.org/10.4232/1.13052.

ISSP Research Group. 2019. International social survey programme: Social networks and social resources - issp 2017. GESIS Data Archive, Cologne. ZA6980 Data file Version 2.0.0, https://doi.org/10.4232/1.13322.

ISSP Research Group. 2020. International social survey programme: Religion iv - issp 2018. GESIS Data Archive, Cologne. ZA7570 Data file Version 2.1.0, https://doi.org/10.4232/1.13629.

ISSP Research Group. 2022. International social survey programme: Social inequality v - issp 2019. GESIS, Cologne. ZA7600 Data file Version 3.0.0, https://doi.org/10.4232/1.14009.

ISSP Research Group. 2023. International social survey programme: Environment iv - issp 2020. GESIS, Cologne. ZA7650 Data file Version 2.0.0, https://doi.org/10.4232/1.14153.

ISSP Research Group. 2024. International social survey programme: Health and health care ii - issp 2021. GESIS, Cologne. ZA8000 Data file Version 2.0.0, https://doi.org/10.4232/5.ZA8000.2.0.0.

Ke Ji, Yixin Lian, Linxu Li, Jingsheng Gao, Weiyuan Li, and Bin Dai. 2025. Enhancing persona consistency for llms' role-playing using persona-aware contrastive learning. *arXiv preprint arXiv:2503.17662*.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts.

Nikitas Karanikolas, Eirini Manga, Nikoletta Samaridi, Eleni Tousidou, and Michael Vassilakopoulos. 2024. Large language models versus natural language understanding and generation. In *Proceedings of the 27th Pan-Hellenic Conference on Progress in Computing and Informatics*, PCI '23, page 278–290, New York, NY, USA. Association for Computing Machinery.

Yan Leng and Yuan Yuan. 2023. Do llm agents exhibit social behavior? *arXiv preprint arXiv:2312.15198*.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36:51991–52008.

Kun Li, Chenwei Dai, Wei Zhou, and Songlin Hu. 2024. Fine-grained behavior simulation with role-playing large language model on social media.

Yuan Li, Lichao Sun, and Yixuan Zhang. 2025. Metaagents: Large language model based agents for decision-making on teaming. *Proc. ACM Hum.-Comput. Interact.*, 9(2).

Yijun Liu, Wu Liu, Xiaoyan Gu, Yong Rui, Xiaodong He, and Yongdong Zhang. 2024. Lmagent: A large-scale multimodal agents society for multi-user simulation.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Comput. Surv.*, 56(2).

Xinyi Mou, Xuanwen Ding, Qi He, Liang Wang, Jingcong Liang, Xinnong Zhang, Libo Sun, Jiayu Lin, Jie Zhou, Xuanjing Huang, and Zhongyu Wei. 2024a. From individual to society: A survey on social simulation driven by large language model-based agents.

Xinyi Mou, Jingcong Liang, Jiayu Lin, Xinnong Zhang, Xiawei Liu, Shiyue Yang, Rong Ye, Lei Chen, Haoyu Kuang, Xuanjing Huang, et al. 2024b. Agentsense: Benchmarking social intelligence of language agents through interactive scenarios. *arXiv preprint arXiv:2410.19346*.

Qian Niu, Junyu Liu, Ziqian Bi, Pohsun Feng, Benji Peng, Keyu Chen, Ming Li, Lawrence KQ Yan, Yichao Zhang, Caitlyn Heqi Yin, Cheng Fei, Tianyang Wang, Yunze Wang, Silin Chen, and Ming Liu. 2024. Large language models and cognitive science: A comprehensive review of similarities, differences, and challenges.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. 2024. Gpt-4o system card.

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*, pages 1–22.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 technical report.

Yuqi Ren, Renren Jin, Tongxuan Zhang, and Deyi Xiong. 2025. Do large language models mirror cognitive language processing? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2988–3001, Abu Dhabi, UAE. Association for Computational Linguistics.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR.

James WA Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, et al. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7):1285–1295.

Hari Subramonyam, Roy Pea, Christopher Pondoc, Maneesh Agrawala, and Colleen Seifert. 2024. Bridging the gulf of envisioning: Cognitive challenges in prompt based interactions with llms. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Chuanneng Sun, Songjun Huang, and Dario Pompili. 2025. Llm-based multi-agent decision-making: Challenges and future directions. *IEEE Robotics and Automation Letters*, 10(6):5681–5688.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil

Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 technical report.

Lei Wang, Jingsen Zhang, Hao Yang, Zhi-Yuan Chen, Jiakai Tang, Zeyu Zhang, Xu Chen, Yankai Lin, Hao Sun, Ruihua Song, Xin Zhao, Jun Xu, Zhicheng Dou, Jun Wang, and Ji-Rong Wen. 2025. User behavior simulation with large language model-based agents. *ACM Trans. Inf. Syst.*, 43(2).

Xinnong Zhang, Jiayu Lin, Xinyi Mou, Shiyue Yang, Xiawei Liu, Libo Sun, Hanjia Lyu, Yihang Yang, Weihong Qi, Yue Chen, Guanying Li, Ling Yan, Yao Hu, Siming Chen, Yu Wang, Xuanjing Huang, Jiebo Luo, Shiping Tang, Libo Wu, Baohua Zhou, and Zhongyu Wei. 2025. Socioverse: A world model for social simulation powered by llm agents and a pool of 10 million real-world users.

A Details of Data Statistics

A.1 Statistics and Analysis

Figure 3 provides a detailed overview of the structural characteristics of questionnaire items in the SocioBench dataset.

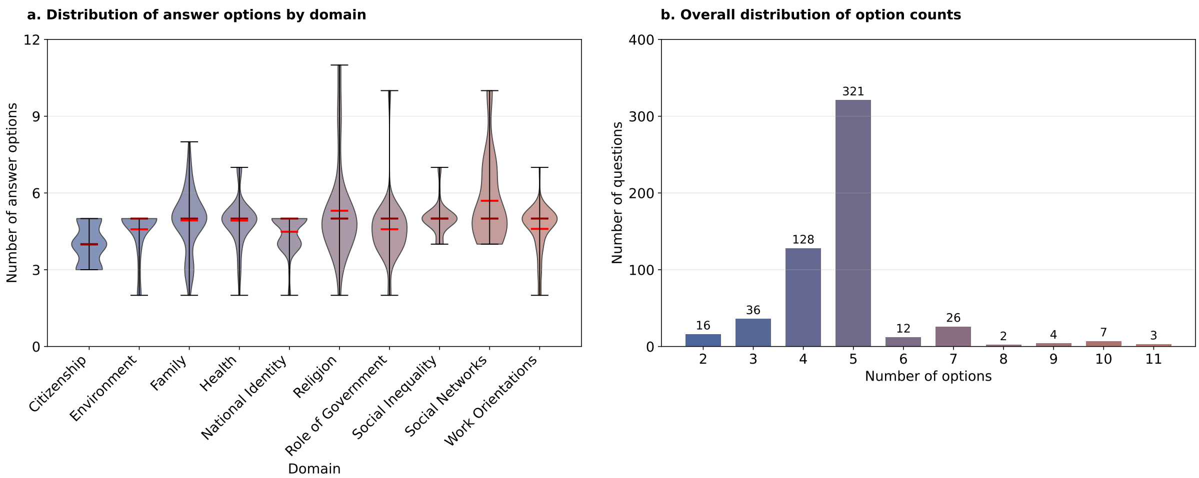


Figure 3: **Question and Answer Option Distribution Analysis across ISSP Survey Domains.** (a) shows the distribution of answer options per question across domains using violin plots. The width of each violin represents the density of questions with that number of options. The red line indicates the mean number of options, while the dark red line shows the median number of options for each domain. The black lines represent the data range (minimum to maximum values). (b) displays the overall distribution of questions grouped by answer option count across the entire dataset, showing how many questions have 2, 3, 4, 5, etc. answer options in total.

A.2 Demographic Information Distribution within the Citizenship Domain

Table 4 to Table 8 show the distribution statistics of some demographic information in the citizenship domain, including gender, country, year of birth, educational background, and religion.

Value	Freq.	Pct.
Austria	16	3.2%
Taiwan, China	16	3.2%
Australia	16	3.2%
Croatia	16	3.2%
Chile	16	3.2%
Lithuania	15	3.0%
Belgium	15	3.0%
Netherlands	15	3.0%
Korea (South)	15	3.0%
Slovakia	15	3.0%
Turkey	15	3.0%

Value	Freq.	Pct.
Venezuela	15	3.0%
United States of America	15	3.0%
Czech Republic	15	3.0%
Germany	15	3.0%
Russia	15	3.0%
Great Britain	15	3.0%
Spain	15	3.0%
Poland	15	3.0%
Georgia	15	3.0%
South Africa	15	3.0%
Norway	15	3.0%

Value	Freq.	Pct.
France	15	3.0%
Japan	15	3.0%
Philippines	15	3.0%
Israel	15	3.0%
India	15	3.0%
Finland	15	3.0%
Switzerland	15	3.0%
Slovenia	15	3.0%
Iceland	15	3.0%
Denmark	15	3.0%
Sweden	15	3.0%

Table 4: Demographic Profile of Citizenship Domain: Country Distribution. Freq. refers to the frequency of occurrence, Pct. refers to the percentage

Value	Freq.	Pct.
1975	15	3.0%
1962	14	2.8%
1961	14	2.8%
1949	13	2.6%
1963	13	2.6%
1965	13	2.6%
1958	13	2.6%
1976	12	2.4%
1981	12	2.4%
1967	12	2.4%
1979	12	2.4%
1964	12	2.4%
1985	12	2.4%
1969	12	2.4%

Value	Freq.	Pct.
1960	12	2.4%
1977	10	2.0%
1951	10	2.0%
1971	10	2.0%
1952	10	2.0%
1992	10	2.0%
1970	9	1.8%
1938	9	1.8%
1974	9	1.8%
1982	9	1.8%
1942	9	1.8%
1950	9	1.8%
1989	8	1.6%
1983	8	1.6%

Value	Freq.	Pct.
1972	8	1.6%
1953	8	1.6%
1954	8	1.6%
1955	8	1.6%
1994	7	1.4%
1947	7	1.4%
1980	7	1.4%
1984	7	1.4%
1973	7	1.4%
1944	7	1.4%
1957	6	1.2%
1986	6	1.2%
1956	6	1.2%
1978	6	1.2%

Value	Freq.	Pct.
1943	6	1.2%
1988	6	1.2%
1946	5	1.0%
1933	5	1.0%
1939	5	1.0%
1948	5	1.0%
1968	5	1.0%
1993	5	1.0%
1987	5	1.0%
1966	4	0.8%
1935	4	0.8%
1959	4	0.8%
1995	4	0.8%
1937	3	0.6%

Value	Freq.	Pct.
1934	3	0.6%
1991	3	0.6%
1936	3	0.6%
1990	3	0.6%
1941	2	0.4%
1998	2	0.4%
1932	2	0.4%
1940	1	0.2%
No answer	1	0.2%
1931	1	0.2%
1925	1	0.2%
1997	1	0.2%
1996	1	0.2%
1945	1	0.2%

Table 5: Demographic Profile of Citizenship Domain: Birth Year Distribution

Value	Freq.	Pct.
Upper secondary (programs that allow entry to university)	122	24.4%
Lower level tertiary, first stage (also technical schools at a tertiary level)	111	22.2%
Lower secondary (secondary completed does not allow entry to university: obligatory school)	106	21.2%
Upper level tertiary (Master, Doctor)	65	13.0%
Post secondary, non-tertiary (other upper secondary programs toward labour market or technical formation)	59	11.8%
Primary school (elementary education)	22	4.4%
No formal education	14	2.8%
No answer	1	0.2%

Table 6: Demographic Profile of Citizenship Domain: Education Level Distribution

Value	Freq.	Pct.
Male	257	51.4%
Female	243	48.6%

Table 7: Demographic Profile of Citizenship Domain: Gender Distribution

Value	Freq.	Pct.
No religion	140	28.0%
Catholic	139	27.8%
Protestant	100	20.0%
Orthodox	26	5.2%
Islamic	25	5.0%
Other Christian	17	3.4%
Buddhist	14	2.8%
Hindu	14	2.8%
Jewish	10	2.0%
Other Asian Religions	5	1.0%
No answer	3	0.6%
Other Religions	3	0.6%
Refused	3	0.6%
Information insufficient	1	0.2%

Table 8: Demographic Profile of Citizenship Domain: Religious Affiliation Distribution

A.3 Data example

Figure 4 and Figure 5 respectively show the questionnaire data, respondent profile data and ground-truth answer data contained in the SocioBench dataset. Figure 4 shows the Q&A data processing for special countries. For example, for question V44, when the respondent's country code is equal to the country code of "special" in the dataset, the corresponding question option in "special" replaces the question option in "answer" and asks the question.


```
{
  "question_id": "V44",
  "question": "Q40 To what extent do you agree or disagree with the following statements? I think most people in [COUNTRY] are
better informed about politics and government than I am.",
  "answer": {
    "1": "Strongly agree",
    "2": "Agree",
    "3": "Neither agree nor disagree",
    "4": "Disagree",
    "5": "Strongly disagree"
  },
  "special": {
    "JP": {
      "1": "I think so",
      "2": "I rather think so",
      "3": "Can't say one way or the other",
      "4": "I rather don't think so",
      "5": "I don't think so"
    },
    "VE": {
      "1": "I agree",
      "2": "I somewhat agree",
      "3": "I neither agree nor disagree",
      "4": "I somewhat disagree",
      "5": "I disagree"
    }
  }
},
```


Figure 4: SocioBench Dataset: Questions and answers in social survey questionnaires

```
[
  {
    "person_id": 10021906,
    "attributes": {
      "Country Prefix ISO 3166": "Lithuania",
      "Sex of Respondent": "Female",
      "Year of birth": "1946",
      "Age of respondent": "69",
      "Education I: years of schooling": "15",
      "Country specific highest completed degree of education: Lithuania": "Vocational (completing basic)",
      "Highest completed education level: Categories for international comparison": "Lower secondary (secondary completed does not
allow entry to university: obligatory school)",
      "Currently, formerly, or never in paid work": "Currently not in paid work, paid work in the past",
      "Hours worked weekly": "NAP (code 2 or 3 in WORK)",
      "Employment relationship": "Employee",
      "Self-employed: how many employees": "NAP (code 1, 2, 4, 0 in EMPREL)",
      "Supervise other employees": "No",
      "Number of other employees supervised": "NAP (code 2, 0 in WRKSUP)",
      "Type of organization, for-profit/ non-profit": "For-profit organization",
      "Type of organization, public/ private": "Public employer",
      "Occupation ISCO/ ILO 2008": "Engineering professionals (excluding electrotechnology)",
      "Main status": "Retired",
      "Living in steady partnership": "Yes, have partner; live in same household",
      "Spouse, partner: currently, formerly or never in paid work": "Currently not in paid work, paid work in the past",
      "Spouse, partner: hours worked weekly": "NAP (code 0, 2 or 3 in SPWORK)",
      "Spouse, partner: employment relationship": "Employee",
      "Spouse, partner: supervise other employees": "No",
      "Spouse, partner: occupation ISCO/ ILO 2008": "Electronics mechanics and servicers",
      "Spouse, partner: main status": "Retired",
      "Trade union membership": "Yes, previously, but not currently",
      "Country specific religious affiliation or denomination: Lithuania": "Orthodox",
……
……
……
    },
    "questions_answer": {
      "v5": 6,
      "v6": 7,
      "v7": 7,
      "v8": 7,
      "v9": 5,
      "v10": 7,
……
……
……
      "v60": 3,
      "v61": 2,
      "v62": 2,
      "v63": 3,
      "v64": 1
     }
  },
```


Figure 5: SocioBench Dataset: respondent demographic information and Ground-truth answers

B Data Curation Details

Figures 6 and 7 show structured Questionnaire QA/Demographic Questionnaire QA examples extracted from the ISSP Variable Report.pdf, in Chinese and English versions, respectively.

Structured Extraction Questionnaire QA/Demographic Questionnaire QA from ISSP Variable Report.pdf

你是一个专业的数据处理专家，请仔细阅读当前pdf，根据我的要求，逐页提取信息，并进行结构化的json输出。
具体包含5点信息：
1.第一点为domain信息，表示当前内容所属的内容缩写，例如："v1"、"C_ALPHAN"、"V9"、"CZ_V65"、"IN_RINC"..........这里只是作为示例，具体内容以当前文档为准，示例中的信息与当前pdf无关，仅供参考；
2.第二点为含义信息，表示domain所指内容，例如："GESIS Data Archive Study Number - 'Citizenship II'"、"Country Prefix ISO 3166"、"Q5 Good citizen: active in social or political associations"、"Q61 Frequency: read political content of a newspaper"、"Country specific personal income: India"等，示例仅作参考；
3.第三点为问题信息，表示调查中具体询问的内容，例如：" GESIS Data Archive Study number ZA6670 for the ISSP 2014 on 'Citizenship II'. Study number of the data set producer and archiving number "、" Sample Prefix ISO 3166 Code - alphanumeric ISO 3166 Country/ Sample Prefix This alphanumerical sample identification variable C_ALPHAN includes country codes that are based on ISO 3166."、" There are different opinions as to what it takes to be a good citizen. As far as you are concerned personally on a scale of 1 to 7, where 1 is not at all important and 7 is very important, how important is it: To be active in social or political associations"、" Before taxes and other deductions, what on average is your own total monthly income? "、"Here are some different forms of political and social action, that people can take. Please indicate, for each one, whether you have done any of these things in the past year, whether you have done it in the more distant past, whether you have not done it but might do it or have not done it and would never, under any circumstances, do it. Attended a political meeting or rally"
- 注意，对于直接用于社会调查访问的内容，不进行提取，例如 (IF DONE BY INTERNET COUNT AS YES)(IF MORE THAN ONE RESPONSE, CODE THE MORE PARTICIPATIVE ONE - THAT IS, THE ONE CLOSER TO THE LEFT END OF THE SCALE.)等
4.第四点为内容信息，数据格式为一组key value，左侧的为option code，表示选项代码。禁止删减输出，所有选项都要输出，包括：NAP, all other countries、Refused、Don't know、No answer等特殊情况；右侧的为option text，表示此选项对应的文本含义，如"6670 GESIS Data Archive Study Number ZA6670"、"AT = Austria"、" 1 1, Not at all important 2 2 3 3 4 4 5 5 6 6 7 7, Very important 8 Can't choose 9 No answer"、" 1 Several times a day "等，你需要逐个结构化为字典格式例如，6670: "GESIS Data Archive Study Number ZA6670"、AT: "Austria"、1:"1, Not at all important"、2:"2"等；
5.第五列，为特殊数据形式，在某些特定的国家编号下，数据需要特殊处理，！！注意："Note:"之中的信息不做任何的提取/处理。例如" Note: / CZ: For-profit organization means limited liability company, private joint stock company, cooperative, profit-seeking state-owned business etc. Non-profit organization means non-profit non-governmental organization, foundation, public benefit corporation, public administration, local administration, public institution like hospitals, public schools, libraries, police, the military."这些信息完全不管。你需要对在选项之中出现如下特殊国家情况，"in Austria (AT): 0 Not available"这样的选项进行处理，需要按照具体的国家格式化为三元组格式，{ "AT": { 0: "Not available" } }、{ "GB-GBN": { 0: "NAP (code 0, 2, 3 in EMPREL" } }，若无特殊选项，输出空白即可。
注意：
1.！！禁止减少输出、省略输出，输出原文英文，禁止修改原始内容的表达，一次性输出完毕当前pdf的全部内容。最终将所有内容输出到一个json中，每一条信息都包含5元组.一次性输出完毕所有页面的信息，禁止不全输出或中途停止。
2.具体内容并非一定与当前pdf相关，上述prompt给出的所有例子禁止直接作为输出，你需要阅读pdf中的内容，在进行输出，必须确保输出内容，直接在pdf中有所对应。我给出你1个输出的示例：

```
{
  "domain": "NEMPLOY",
  "meaning": "Self-employed: how many employees",
  "question": "If self-employed with employees, how many employees do/did you have, not counting yourself?",
  "content": {
    "0": "NAP (code 1, 2, 4, 0 in EMPREL)",
    "1": "1 employee",
    "9995": "9995 employees or more",
    "9998": "Don't know",
    "9999": "No answer"
  },
  "special": {
    "NL": {
      "4": "2-5 employees",
      "9": "6-11 employees",
      "19": "12-25 employees",
      "30": "More than 25 employees"
    },
    "US": {
      "97": "97 employees or more"
    }
  }
}
```

Figure 6: Structured Extraction Questionnaire QA/Demographic Questionnaire QA from ISSP Variable Report.pdf (Chinese)

Structured Extraction Questionnaire QA/Demographic Questionnaire QA from ISSP Variable Report.pdf

You are a professional data processing expert. Please carefully read the current PDF and, according to my requirements, extract information page by page and output it in a structured JSON format.

Specifically, the output should include the following five pieces of information:

1. Domain Information: Indicates the abbreviation of the current content's domain, such as "v1", "C_ALPHAN", "V9", "CZ_V65", "IN_RINC", etc. These are just examples; the actual content should be based on the current document. The examples provided are for reference only and are not related to the current PDF.

2. Meaning Information: Represents the meaning of the domain, for example: "GESIS Data Archive Study Number - 'Citizenship II'", "Country Prefix ISO 3166", "Q5 Good citizen: active in social or political associations", "Q61 Frequency: read political content of a newspaper", "Country specific personal income: India", etc. These examples are for reference only.

3. Question Information: Indicates the specific question asked in the survey, such as: "GESIS Data Archive Study number ZA6670 for the ISSP 2014 on 'Citizenship II'. Study number of the data set producer and archiving number", "Sample Prefix ISO 3166 Code - alphanumeric ISO 3166 Country/ Sample Prefix This alphanumerical sample identification variable C_ALPHAN includes country codes that are based on ISO 3166.", "There are different opinions as to what it takes to be a good citizen. As far as you are concerned personally on a scale of 1 to 7, where 1 is not at all important and 7 is very important, how important is it: To be active in social or political associations", "Before taxes and other deductions, what on average is your own total monthly income?", "Here are some different forms of political and social action that people can take. Please indicate, for each one, whether you have done any of these things in the past year, whether you have done it in the more distant past, whether you have not done it but might do it, or have not done it and would never, under any circumstances, do it. Attended a political meeting or rally", etc.

Note: Do not extract content that is directly used for social survey visits, such as "(IF DONE BY INTERNET COUNT AS YES)(IF MORE THAN ONE RESPONSE, CODE THE MORE PARTICIPATIVE ONE - THAT IS, THE ONE CLOSER TO THE LEFT END OF THE SCALE.)", etc.

4. Content Information: The data format should be a set of key-value pairs, where the left side is the option code, representing the option code, and the right side is the option text, representing the textual meaning of the option, such as "6670 GESIS Data Archive Study Number ZA6670", "AT = Austria", "1 1, Not at all important 2 2 3 3 4 4 5 5 6 6 7 7, Very important 8 Can't choose 9 No answer", "1 Several times a day", etc. You need to structure each as a dictionary format, for example, 6670: "GESIS Data Archive Study Number ZA6670", AT: "Austria", 1: "1, Not at all important", 2: "2", etc.

5. Special Data Format: In certain specific country codes, data requires special handling. **Note**: Information within "Note:" should not be extracted or processed in any way. For example, "Note: / CZ: For-profit organization means limited liability company, private joint stock company, cooperative, profit-seeking state-owned business, etc. Non-profit organization means non-profit non-governmental organization, foundation, public benefit corporation, public administration, local administration, public institution like hospitals, public schools, libraries, police, the military." This information should be completely ignored.

You need to process special country cases that appear in the options, such as "in Austria (AT): 0 Not available". These options should be handled and formatted into a triple format, for example, { "AT": { 0: "Not available" } }, { "GB-GBN": { 0: "NAP (code 0, 2, 3 in EMPREL" } }. If there are no special options, leave this blank.

Attention:
1. Do not reduce or omit any output**; output the original English text without modifying the original expressions. Output all the content of the current PDF at once. Each piece of information should include a 5-tuple. Do not output partially or stop midway.

2. The specific content may not necessarily be related to the current PDF. The examples provided in the above prompt are not to be directly used as output. You need to read the content in the PDF and then output accordingly. Ensure that the output content directly corresponds to the content in the PDF.

I will provide you with an example of the output:

```json
{
  "domain": "NEMPLOY",
  "meaning": "Self-employed: how many employees",
  "question": "If self-employed with employees, how many employees do/did you have, not counting yourself?",
  "content": {
    "0": "NAP (code 1, 2, 4, 0 in EMPREL)",
    "1": "1 employee",
    "9995": "9995 employees or more",
    "9998": "Don't know",
    "9999": "No answer"
  },
  "special": {
    "NL": {
      "4": "2-5 employees",
      "9": "6-11 employees",
      "19": "12-25 employees",
      "30": "More than 25 employees"
    },
    "US": {
      "97": "97 employees or more"
    }
  }
}
```

Figure 7: Structured Extraction Questionnaire QA/Demographic Questionnaire QA from ISSP Variable Report.pdf (English)

C LLMs Role-playing Prompt template

Figure 8 shows the prompt templates passed to the LLMs along with examples.

```
### Instruction:
You are participating in the International Social Survey Programme. Assume the role of a real individual with the following personal informa-
tion. Fully immerse yourself in this persona and answer the question truthfully, based solely on the provided personal information.

### Personal Information:
{attributes}

e.g.
Country Prefix ISO 3166: Australia;
Sex of Respondent: Female; Year of birth: 1984;
Age of respondent: 31; Education I: years of schooling: 11;
Country specific highest completed degree of education: Australia: Completed High School to Year 10;
Highest completed education level: Categories for international comparison: Lower secondary (secondary completed does not allow entry to uni-
versity: obligatory school);
Currently, formerly, or never in paid work: Currently in paid work;
Hours worked weekly: 40;
Employment relationship: Employee;
Supervise other employees: Yes;
Number of other employees supervised: 12;
Type of organization, for-profit/ non-profit: For-profit organization;
Type of organization, public/ private: Private employer;
Occupation ISCO/ ILO 2008: No answer;
Main status: In paid work;
Living in steady partnership: Yes, have partner; live in same household;
Spouse, partner-currently, formerly or never in paid work: Currently in paid work;
Spouse, partner-hours worked weekly: 48;
Spouse, partner-employment relationship: Employee; Spouse, partner: supervise other employees: No;
Spouse, partner-occupation ISCO/ ILO 2008: Advertising and marketing professionals;
Spouse, partner-main status: In paid work; Trade union membership: No, never;
Country specific religious affiliation or denomination: Australia: No religion;
Groups of religious affiliations (derived from nat_RELIG): No religion;
Attendance of religious services: Never;
Top-Bottom self-placement: No answer;
Did respondent vote in last general election: Yes;
Country specific party voted for in last general election-Australia: Australian Labor Party - ALP;
Party voted for in last general election: left-right (derived from nat_PRTY): Left, center left;
Country specific ethnic group 1: Australia: AU born: + Both parents also AU born;
How many children in household: children between [school age] and 17 years of age: No children;
How many toddlers in household: children up to [school age -1] years: No toddlers;
How many persons in household: Two persons;
Australia: Country specific personal income: 5200;
Australia: Country specific household income: 12000;
Legal partnership status: Married; Father's country of birth: Australia;
Mother's country of birth: Australia;
Place of living: urban - rural: The suburbs or outskirts of a big city;
Australia: Country specific region: South Australia;
person_id: 10001310.
### Question:
{question}

e.g.
Q1 There are different opinions as to what it takes to be a good citizen. As far as you are concerned personally on a scale of 1 to 7, where
1 is  not at all important and 7 is very important, how important is it: Always to vote in elections

### Options:
{options}

e.g.
4: 4; 5: 5; 7: 7, Very important; 1: 1, Not at all important; 2: 2; 6: 6; 3: 3

### Please strictly follow the following json format output:
```json
{{{{
  "reason": "",
  "option": ""
}}}}

```
### Requirements:
1. Please answer the questions based on your personal information only and give a detailed and complete justification, which requires a 6-10
sentence response.
2. 2. Please choose the option that best suits you from the ### Options given, and respond with the number only. For example: #### Op-
tions contains: {{"1": "1, Not at all important", "2": "02"}}, you can choose "1" or "2", but do not choose "1, Not at all important" or "02".
```

Figure 8: Prompt Template for LLMs Role-playing Respondents in Social Survey Scenarios

D Comparison of Qwen3-32B With and Without "Think" Mode

Figure 9 shows the Qwen3-32B response comparison on the same question.

Figure 9: Qwen3-32B response comparison on the same question

E Comparison of Qwen3-32B With and Without "Reason" in prompt

Figure 10 shows the Qwen3-32B response comparison on the same question.

Question: There are different opinions as to what it takes to be a good citizen. As far as you are concerned personally on a scale of 1 to 3, where 1 is not at all important and 3 is very important, how important is it: Always to vote in elections.
Options: 3: 3, Very important; 2: 2; 1: 1, Not at all important.

1 Qwen3-32B w/ reason

"llm_response": "<think>\nOkay, let's tackle this question. The user is asking about the importance of always voting in elections on a scale of 1 to 3. The options are 1 (not at all important), 2, or 3 (very important). \n\nFirst, I need to look at the personal information provided. The respondent is a 69-year-old Lithuanian woman with a vocational education. She's retired, lives in a town in Vilnius, and is married to someone who's also retired. She's Orthodox and voted in the last election for the Electoral Action of Lithuanian Poles - LLRA. She's part of a household with three people, no children.\n\nNow, the question is about her personal belief on the importance of voting. Since she did vote in the last general election, that suggests she values participation in the democratic process. Her party is on the left-right spectrum as \"Other,\" which might indicate she's not strictly aligned with major parties but still engages politically. Her education level is lower secondary, which might influence her political awareness, but she's active enough to vote. \n\nHer religious affiliation as Orthodox could play a role, but there's no direct link to voting behavior mentioned. However, her involvement in a political party and voting history imply she sees voting as important. The fact that she's part of a trade union (previously) might also mean she's more engaged with societal issues. \n\nConsidering her age and the fact that she's retired, she might have more time to engage in civic duties. The question is about the importance of always voting. Since she voted in the last election, it's likely she views it as important. But the scale is 1-3, with 3 being very important. Given that she's a voter and her party affiliation, she might rate it as 3. However, the options don't have a 3 labeled as \"Very important\" in the options provided. Wait, looking back, the options are 2: 2, 3: 3, Very important, and 1: 1, Not at all important. So the correct answer would be 3. \n\nBut wait, the user's personal info says she voted in the last election, which suggests she believes in voting. Therefore, she would rate it as important. Since the options include 3 as \"Very important,\" that's the most fitting. Even though she's retired, her active participation in the last election indicates she values it. So the answer should be 3.\n</think>\n\n{\n \"reason\": \"As a retired individual who actively voted in the last general election for a specific political party, I believe voting is a fundamental civic duty. My participation in the electoral process reflects a commitment to shaping societal decisions. Although I am not currently employed, my historical engagement with political systems (e.g., trade union membership) underscores the importance of civic responsibility. Voting aligns with my values of contributing to collective governance, even if my current status as a non-worker shifts priorities. My Orthodox religious background may emphasize community and accountability, further reinforcing the significance of electoral participation. The fact that I consistently voted in the past suggests a strong belief in its importance, warranting the highest rating on the scale.\",\n \"option\": \"3\"\n}",

2 Qwen3-32B w/o reason

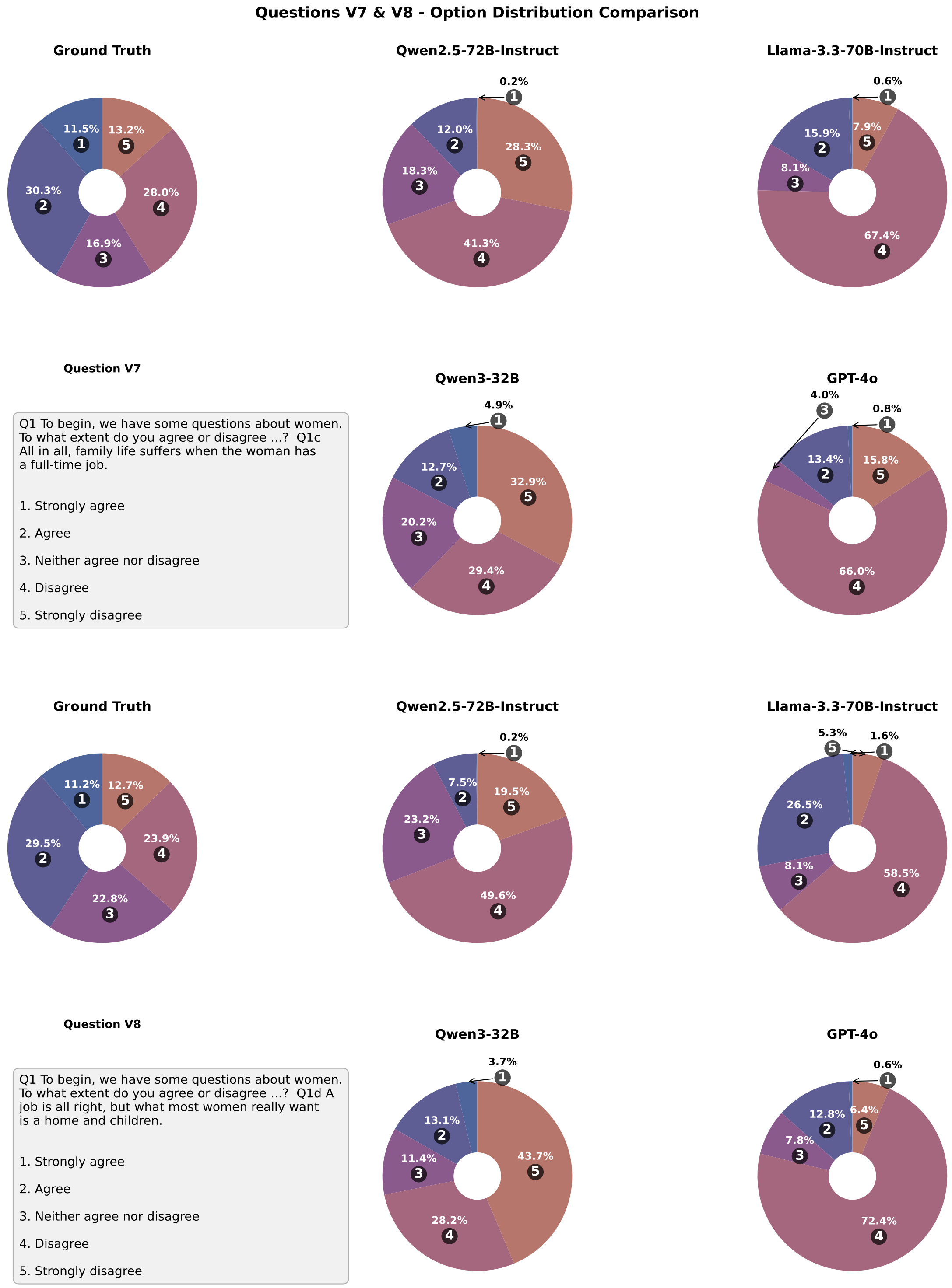
"llm_response": "<think>\nOkay, let's see. The question is about how important it is to always vote in elections, on a scale of 1 to 3. The options are 1 (not at all important), 2, or 3 (very important). \n\nLooking at the personal information provided: the respondent is a 69-year-old woman from Lithuania, retired, with a lower secondary education. She's in a steady partnership, both she and her partner are retired. She's Orthodox, attends religious services once a month, and self-places on the top-bottom scale at 05 (which I think is neutral). She did vote in the last general election and supported the Electoral Action of Lithuanian Poles - LLRA, which is categorized as 'Other' in left-right terms. \n\nShe's part of the Russian ethnic group in Lithuania. Since she did vote in the last election, that might indicate she sees some importance in voting. Also, considering her age and being retired, she might have a stronger sense of civic duty. However, her education level is lower secondary, which might influence her perspective. But since she actually voted, it's likely she believes it's at least somewhat important. The options are 1, 2, or 3. Since she voted, she probably wouldn't rate it as 1. Between 2 and 3. Given that she's part of a political party that's 'Other' and not strictly left or right, maybe she values voting as important but not extremely. However, the fact that she did vote might push her towards 3. But I need to consider her personal info. She's in a town, Vilnius region. Maybe in Lithuania, voting is seen as important. Also, being part of a minority (Russian ethnic group) might make her feel that voting is a way to have her voice heard. So, she might rate it as very important (3). Alternatively, maybe she's more moderate. But since she actually voted, and the question is about personal importance, it's more likely she thinks it's very important.\n</think>\n\n```json\n{\n \"option\": \"3\"\n}\n```",

Figure 10: Qwen3-32B response comparison on the same question

F Example comparison of the option distribution for real respondents and LLM-generated responses.

Focusing on the *Family* and *Health and Health Care* domains, we conducted a further analysis comparing real respondents with four representative models—Qwen2.5-72B-Instruct, Qwen3-32B, Llama-3.3-70B-Instruct, and GPT-4o—by sampling ten questions and examining the response-option distributions.

As shown in Figure 11 and Figure 12, although the ground-truth results exhibit skewed distributions (i.e., options are concentrated in several categories), the LLM-generated responses make this skewness more pronounced, with Llama-3.3-70B-Instruct showing the most marked concentration. Conversely, we observe that Qwen3-32B tends to produce more uniform option distributions.


Questions V7 & V8 - Option Distribution Comparison
Ground Truth
11.5% 1
13.2% 5
28.0% 4
16.9% 3
30.3% 2
Qwen2.5-72B-Instruct
0.2% 1
12.0% 2
28.3% 5
18.3% 3
41.3% 4
Llama-3.3-70B-Instruct
0.6% 1
15.9% 2
7.9% 5
8.1% 3
67.4% 4
Question V7
Q1 To begin, we have some questions about women. To what extent do you agree or disagree ...? Q1c All in all, family life suffers when the woman has a full-time job.
1. Strongly agree
2. Agree
3. Neither agree nor disagree
4. Disagree
5. Strongly disagree
Qwen3-32B
4.9% 1
12.7% 2
32.9% 5
20.2% 3
29.4% 4
GPT-4o
4.0% 3
0.8% 1
13.4% 2
15.8% 5
66.0% 4
Ground Truth
11.2% 1
12.7% 5
29.5% 2
23.9% 4
22.8% 3
Qwen2.5-72B-Instruct
0.2% 1
7.5% 2
19.5% 5
23.2% 3
49.6% 4
Llama-3.3-70B-Instruct
5.3% 5
1.6% 1
26.5% 2
8.1% 3
58.5% 4
Question V8
Q1 To begin, we have some questions about women. To what extent do you agree or disagree ...? Q1d A job is all right, but what most women really want is a home and children.
1. Strongly agree
2. Agree
3. Neither agree nor disagree
4. Disagree
5. Strongly disagree
Qwen3-32B
3.7% 1
13.1% 2
11.4% 3
43.7% 5
28.2% 4
GPT-4o
0.6% 1
12.8% 2
6.4% 5
7.8% 3
72.4% 4

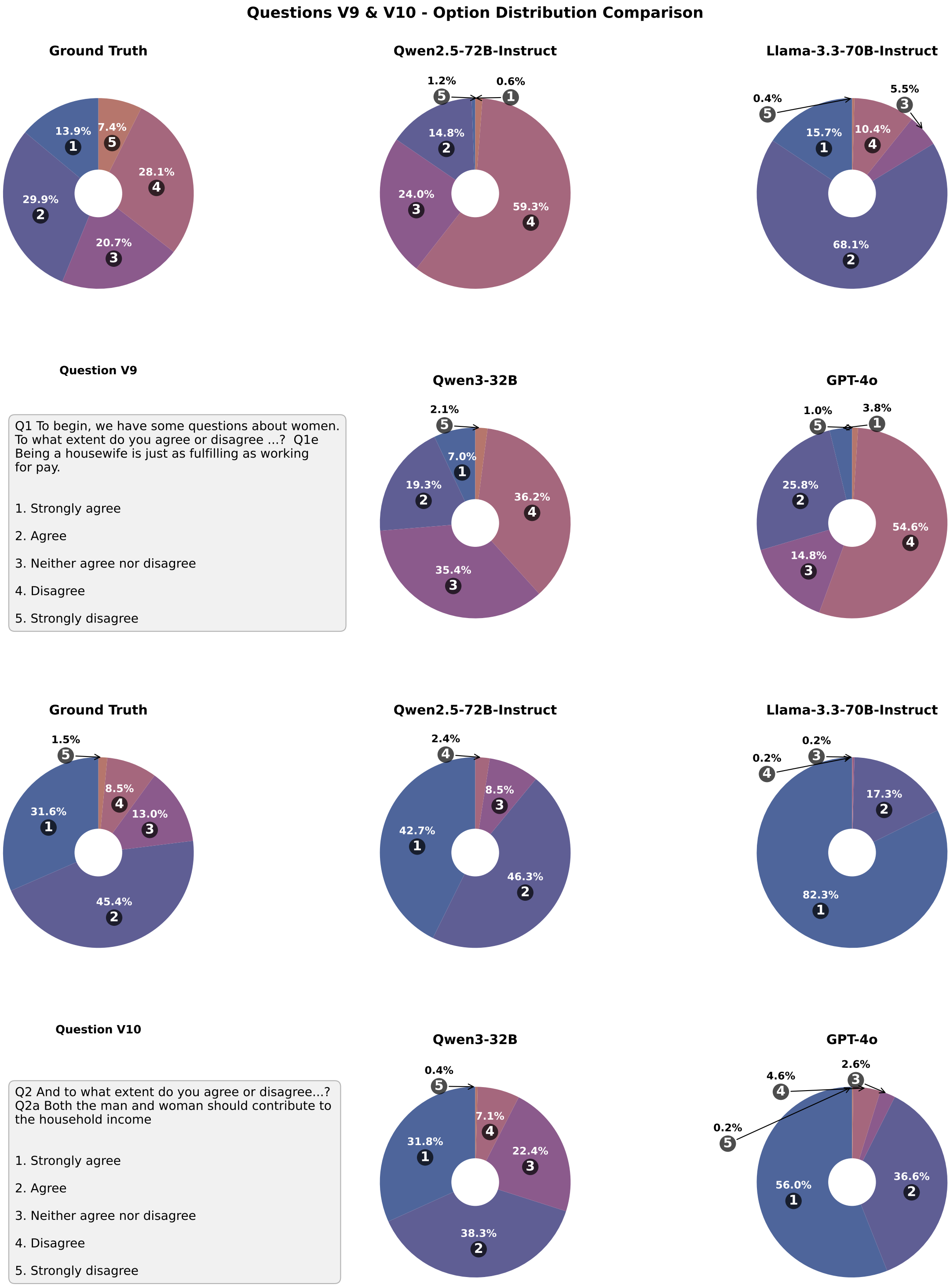
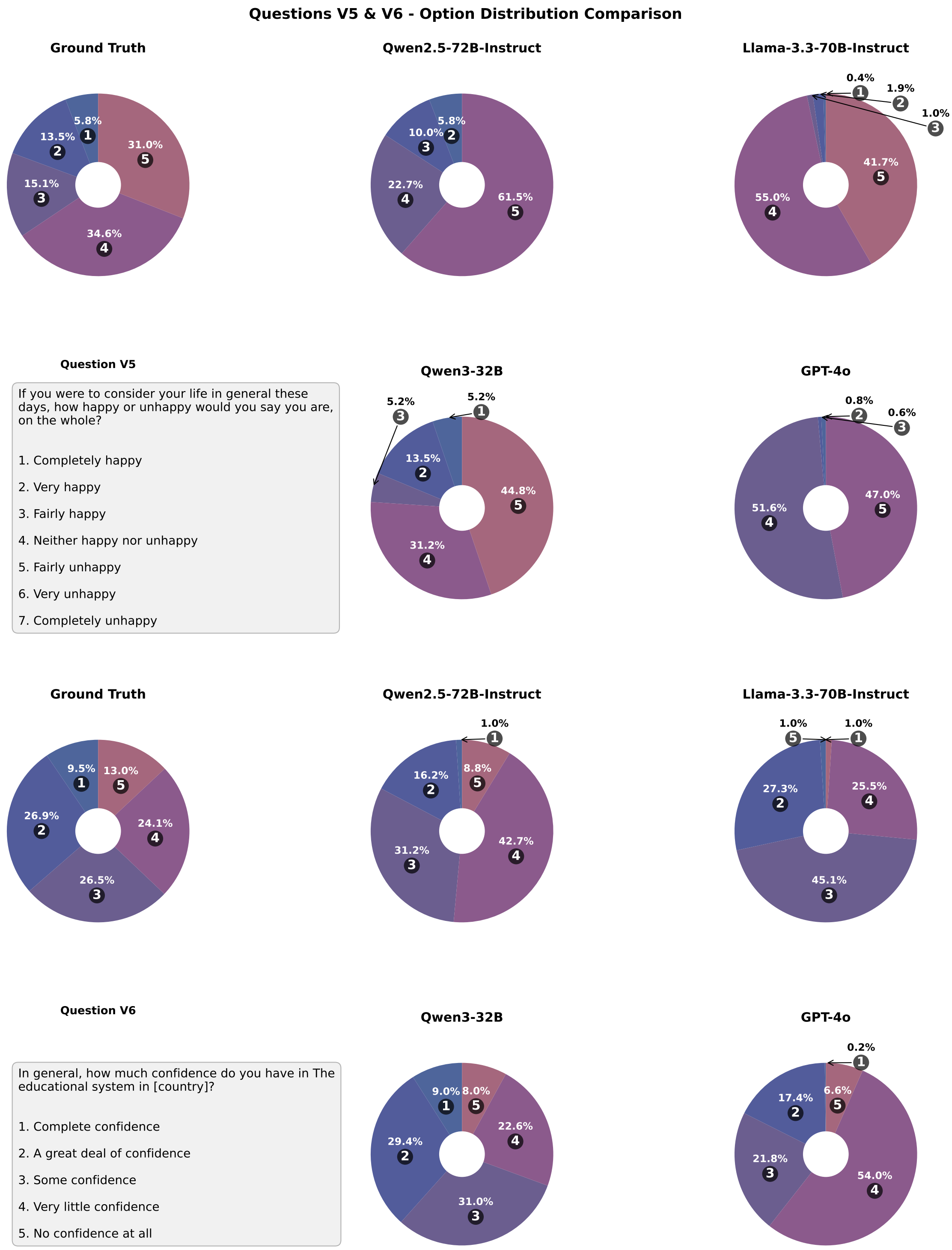
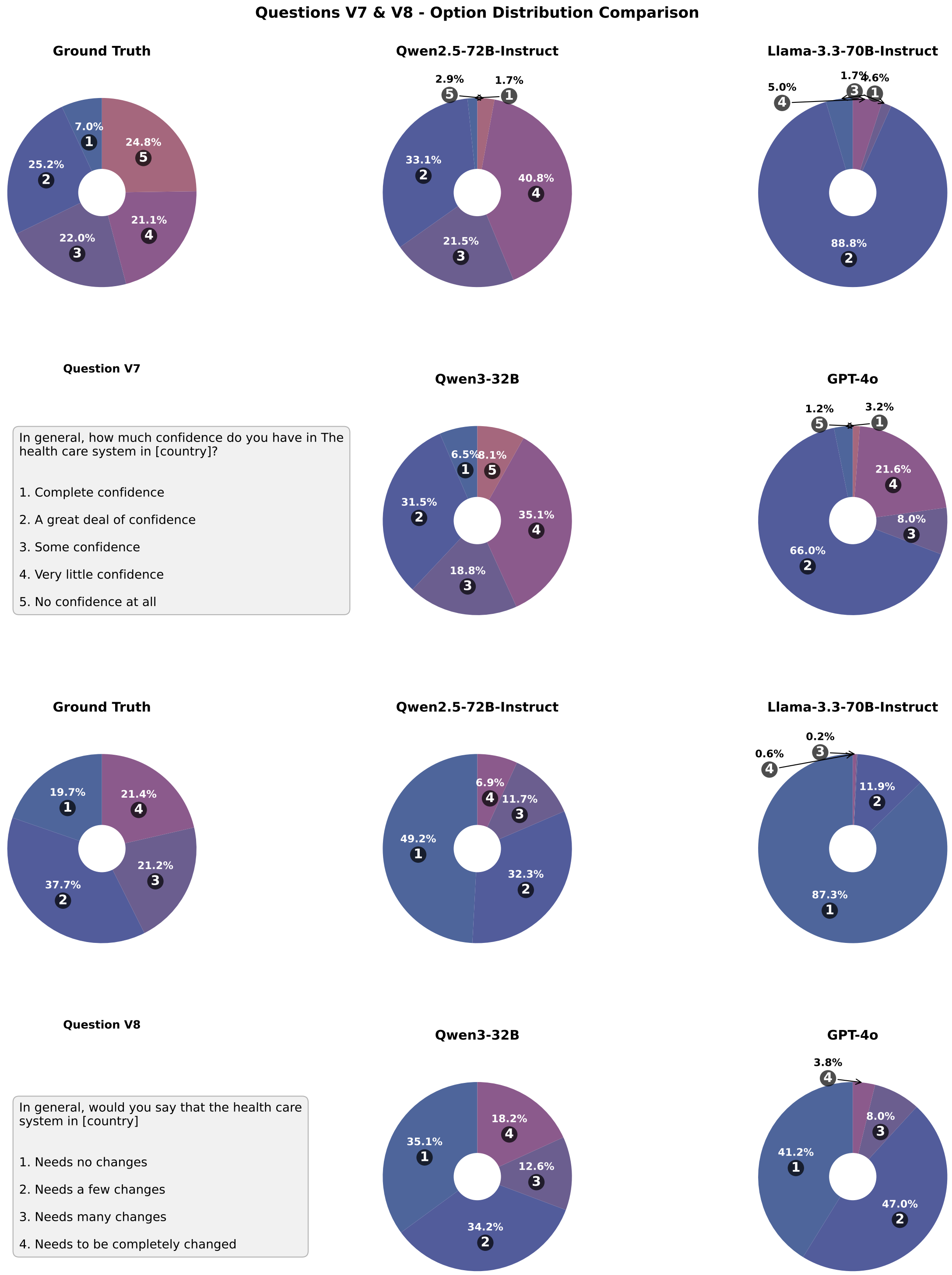


Figure 11: Comparison of the option distribution in the family domain


Questions V5 & V6 - Option Distribution Comparison
Ground Truth
5.8% 1
13.5% 2
15.1% 3
34.6% 4
31.0% 5
Qwen2.5-72B-Instruct
5.8% 2
10.0% 3
22.7% 4
61.5% 5
Llama-3.3-70B-Instruct
0.4% 1
1.9% 2
1.0% 3
55.0% 4
41.7% 5
Question V5
If you were to consider your life in general these days, how happy or unhappy would you say you are, on the whole?
1. Completely happy
2. Very happy
3. Fairly happy
4. Neither happy nor unhappy
5. Fairly unhappy
6. Very unhappy
7. Completely unhappy
Qwen3-32B
5.2% 1
13.5% 2
5.2% 3
31.2% 4
44.8% 5
GPT-4o
0.8% 2
0.6% 3
51.6% 4
47.0% 5
Ground Truth
9.5% 1
26.9% 2
26.5% 3
24.1% 4
13.0% 5
Qwen2.5-72B-Instruct
1.0% 1
16.2% 2
31.2% 3
42.7% 4
8.8% 5
Llama-3.3-70B-Instruct
1.0% 1
27.3% 2
45.1% 3
25.5% 4
1.0% 5
Question V6
In general, how much confidence do you have in The educational system in [country]?
1. Complete confidence
2. A great deal of confidence
3. Some confidence
4. Very little confidence
5. No confidence at all
Qwen3-32B
9.0% 1
29.4% 2
31.0% 3
22.6% 4
8.0% 5
GPT-4o
0.2% 1
17.4% 2
21.8% 3
54.0% 4
6.6% 5


Questions V7 & V8 - Option Distribution Comparison
Ground Truth
7.0% 1
24.8% 5
25.2% 2
21.1% 4
22.0% 3
Qwen2.5-72B-Instruct
2.9% 5
1.7% 1
33.1% 2
40.8% 4
21.5% 3
Llama-3.3-70B-Instruct
5.0% 4
1.7% 3
0.6% 1
88.8% 2
Question V7
In general, how much confidence do you have in The health care system in [country]?
1. Complete confidence
2. A great deal of confidence
3. Some confidence
4. Very little confidence
5. No confidence at all
Qwen3-32B
6.5% 1
8.1% 5
31.5% 2
35.1% 4
18.8% 3
GPT-4o
1.2% 5
3.2% 1
21.6% 4
8.0% 3
66.0% 2
Ground Truth
19.7% 1
21.4% 4
21.2% 3
37.7% 2
Qwen2.5-72B-Instruct
6.9% 4
11.7% 3
49.2% 1
32.3% 2
Llama-3.3-70B-Instruct
0.2% 3
0.6% 4
11.9% 2
87.3% 1
Question V8
In general, would you say that the health care system in [country]
1. Needs no changes
2. Needs a few changes
3. Needs many changes
4. Needs to be completely changed
Qwen3-32B
18.2% 4
35.1% 1
12.6% 3
34.2% 2
GPT-4o
3.8% 4
8.0% 3
41.2% 1
47.0% 2

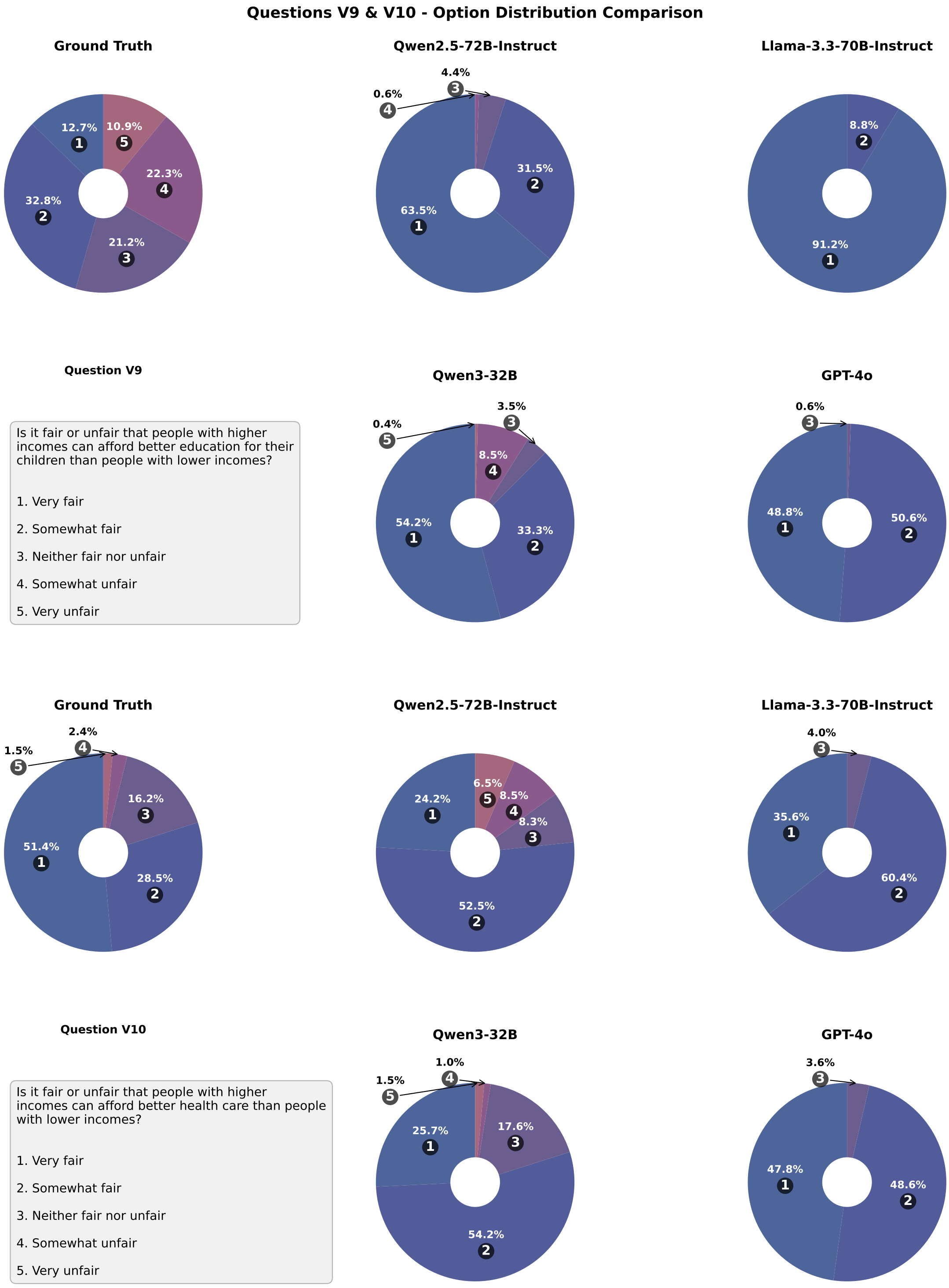


Figure 12: Comparison of the option distribution in the health domain

G Supplementary experimental results and findings

Data Sampling Ratios Comparison. For the experimental result of dataset sampling ratios, please refer to the Table 9.

	n = 5000	n = 10000	n = 20000
Citizen	40.43	40.15	40.07
Enviro	32.11	32.33	32.08
Family	31.89	32.82	33.19
Health	32.21	32.55	32.47
Nat.Ident	33.37	33.37	33.27
Religion	36.99	37.17	36.83
R.Gov	35.27	34.85	34.87
S.Ineq	31.47	31.03	31.48
S.Net	34.99	34.43	35.02
Work	33.39	34.10	33.94
Avg.	34.21	34.28	34.32

Table 9: Results of Llama-3.1-8B-Instruct under different sampling ratios (n denotes the number of respondents under 10 domains).

How does requiring a reason in responses affect LLMs' behavioral simulation? To analyze how providing reasons impacts the evaluation, we conducted experiments on Qwen3-8B and Qwen3-32B, comparing two response strategies: *Option-only* vs. *Reason & Option*. The results indicate that including reasons has a minor effect on performance. In fact, it leads to a slight decrease in accuracy, as detailed in Table 10. We analyse that this may be due to the cognitive overhead or response biases, which can interfere with the model's intrinsic decision-making process. An output example can be found in Appendix E.

	8B w/ R	B w/o R	32B w/ R	32B w/o R
Citizen	40.28	39.96	43.60	44.18
Enviro	32.70	32.64	34.12	34.78
Family	33.07	33.61	34.53	35.37
Health	33.98	34.79	33.53	34.21
Nat.Ident	33.12	34.45	32.64	34.79
Religion	37.58	37.62	38.90	39.52
R.Gov	34.65	34.50	35.52	35.89
S.Ineq	30.83	30.81	33.16	33.14
S.Net	34.38	34.71	35.31	36.54
Work	34.20	35.55	35.25	35.18
Avg.	34.48	**34.87**	35.66	**36.36**

Table 10: Results of Qwen3 models with/without reason in response (R indicates the reason why the LLM selected this option when responding.

How thinking modes shape LLMs' behavioral simulation? For the experimental result of how thinking and reasoning processes affect behavioral simulation in social survey scenarios, please refer to the Table 11.

Comparison Across Survey Rounds. Because the ISSP determines its annual sociological topics through general meetings and typically fields one survey per domain each year, we conducted additional, extensive experiments to compare how survey rounds from different years within the same domain affect benchmark results. Using Llama-3.3-70B-Instruct, we performed experiments for *Environment*, *Health and Healthcare*, *National Identity, Religion*, *Role of Government*, *Social Inequality*, and *Work Orientations*. By contrast, for *Citizenship*, *Family and Changing Gender Roles*, and *Social Networks*, limitations imposed by the data format of the Variable Reports files prevented us from extracting fully structured datasets; therefore, we did not carry out further experiments on these domains, see Table 12.

Across the seven domains with two waves, temporal changes remain modest and bidirectional: *Religion* (+1.95 pp, 2008→2018), *Role of Government* (+1.28 pp, 2006→2016), and *Environment* (+1.28 pp,

	8B w/ T	8B w/o T	32B w/ T	32B w/o T
Citizen	40.28	42.34	43.60	43.52
Enviro	32.70	32.66	34.12	32.63
Family	33.07	30.36	34.53	32.05
Health	33.98	32.23	33.53	33.52
Nat.Ident	33.12	33.59	32.64	31.86
Religion	37.58	37.52	38.90	37.90
R.Gov	34.65	32.94	35.52	35.31
S.Ineq	30.83	30.78	33.16	32.15
S.Net	34.38	33.03	35.31	35.52
Work	34.20	34.25	35.25	33.27
Avg.	**34.48**	33.97	**35.66**	34.77

Table 11: Results of Qwen3 Models With/Without Think Mode (T denotes the think mode; 8B and 32B denote Qwen3–8B and Qwen3–32B, respectively).

2010→2020) show small improvements, while *Work Orientations* (-4.90 pp, 2005→2015), *Health and Healthcare* (-2.48 pp, 2011→2021), *National Identity* (-1.22 pp, 2003→2013), and *Social Inequality* (-1.36 pp, 2009→2019) decline. Averaged across these pairs, the later wave's accuracy is slightly lower by 0.78 pp than the earlier one (37.90% vs. 38.68%), indicating no systematic drift over time.

The benchmark (bold) years used in SocioBench yield an average accuracy of 37.90% (SD=1.90; range 35.73–41.26). The strongest results occur in *Religion* (41.26%) and *Role of Government* (39.19%). A similar pattern is observed in the earlier, non-benchmark waves, which exhibit a comparable mean accuracy of 38.68% (SD=2.54; range 34.69–43.70), with *Work Orientations* (43.70%) and *National Identity* (39.41%) as the top performers. While temporal deltas show some variation—with *Work Orientations* decreasing by 4.90 pp and *Religion* increasing by 1.95 pp over their respective decade spans—most changes remain minor. This suggests that performance is driven more by domain-specific structure than by survey rounds.

As observed from Figure 13, within the same domain, the accuracy between the two survey rounds is highly consistent across continents. For instance, in the *Environment*, performance in the first round is uniformly lower than in the second round for all continents. Conversely, in the *Health and Healthcare* domain, the first round consistently outperforms the second across all continents. This indicates that while accuracy is influenced by the domain and the specific survey round, the benchmark performance demonstrates coordination and consistency across different continents.

Domain	Year	Accuracy
Enviro	2010	34.69
	2020	35.97
Health	2011	38.64
	2021	36.16
Nat.Ident	2003	39.41
	2013	38.19
Religion	2008	39.31
	2018	41.26
R.Gov	2006	37.91
	2016	39.19
S.Ineq	2009	37.09
	2019	35.73
Work	2005	43.70
	2015	38.80

Table 12: Comparison of benchmark accuracy across survey rounds. Years set in bold correspond to the data years used in the SocioBench dataset, whereas years in regular (non-bold) type denote supplementary comparison waves.

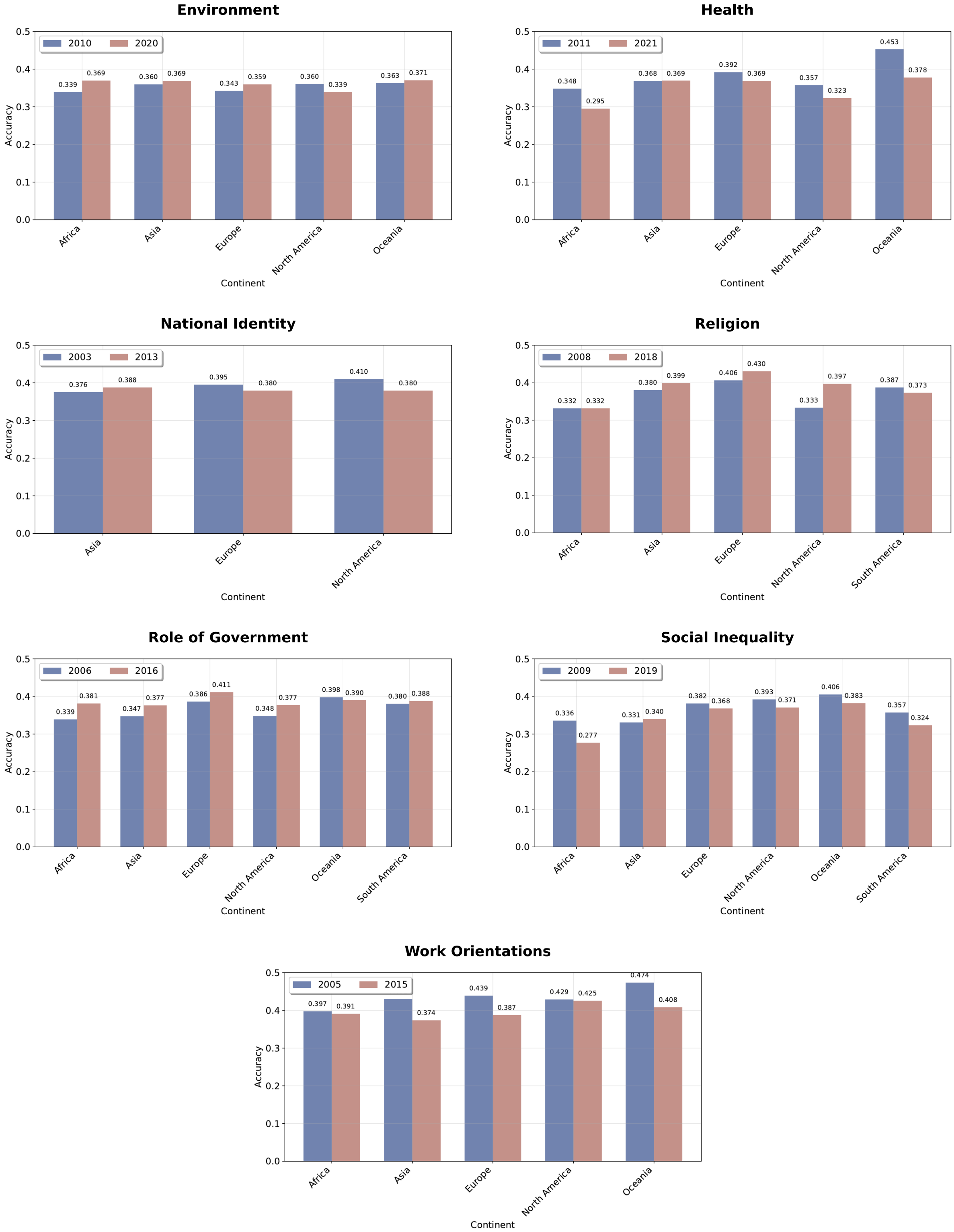


Figure 13: Comparison of benchmark accuracy across different continents in the two survey rounds.

H Subgroup analysis: Biases Across Demographic Information

For the results of subgroup analyses by gender and age, please refer to the Figure 14 and Figure 15.

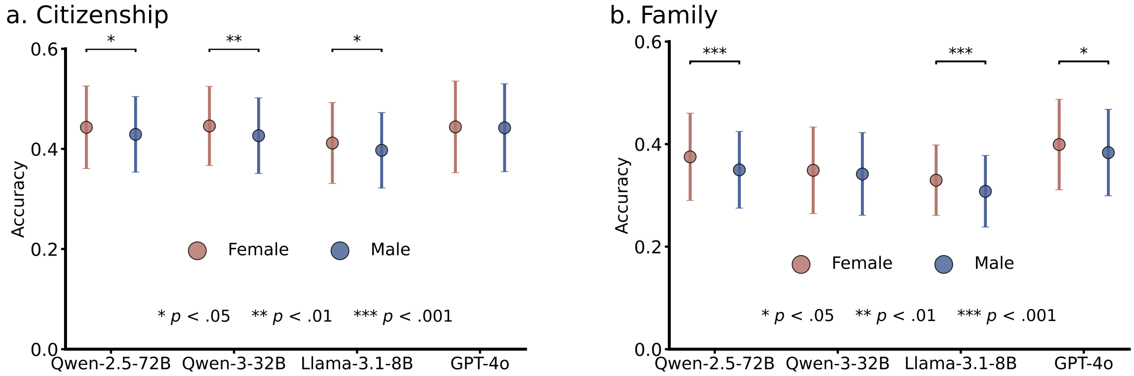


Figure 14: Experimental Results and Significance Analysis of Representative LLMs in the Cross-Gender Subgroup.

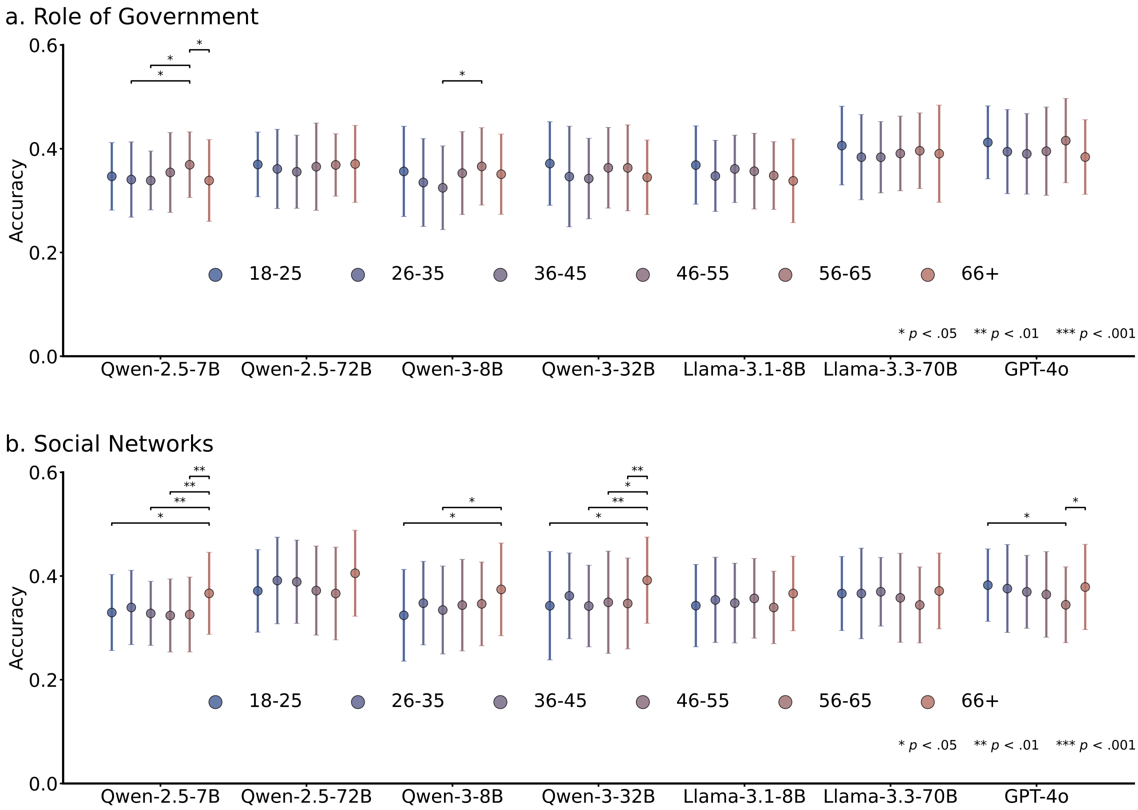


Figure 15: Experimental Results and Significance Analysis of Representative LLMs in the Cross-Age Subgroup.

I Detailed Accuracy by Demographic Variables

This appendix presents the detailed model accuracy results across different demographic subgroups, broken down by domain and variable.

Table 13: Mean Accuracy (% ± SD) of Each Model across Regions for All Domains.

Domain	Continent	Qwen2.5-7B	Qwen2.5-72B	Qwen3-8B	Qwen3-32B	Llama-3.1-8B	Llama-3.3-70B	GPT-4o
Citizenship	Africa	41.8 ± 7.1	39.9 ± 7.8	41.4 ± 7.6	42.1 ± 8.5	38.1 ± 8.5	41.6 ± 7.8	40.6 ± 7.9
	Asia	37.8 ± 10.1	39.8 ± 8.9	39.4 ± 8.4	41.1 ± 9.4	37.2 ± 9.5	39.4 ± 9.6	41.5 ± 10.3
	Europe	41.2 ± 8.5	44.6 ± 7.8	40.4 ± 8.1	44.1 ± 7.5	41.0 ± 7.3	44.9 ± 8.5	45.0 ± 8.6
	North America	37.6 ± 7.6	42.6 ± 7.4	37.9 ± 8.6	40.1 ± 5.8	38.4 ± 6.2	44.0 ± 7.9	46.4 ± 10.8
	Oceania	44.0 ± 7.3	46.0 ± 5.6	42.0 ± 8.5	44.6 ± 4.5	44.3 ± 7.6	49.8 ± 5.9	48.2 ± 5.4
	South America	45.0 ± 6.0	43.3 ± 5.0	41.4 ± 5.7	46.3 ± 6.1	42.3 ± 8.5	44.1 ± 7.1	43.2 ± 7.2
Environment	Africa	28.1 ± 6.8	31.1 ± 9.0	35.4 ± 8.1	30.0 ± 6.6	29.9 ± 6.9	36.4 ± 7.5	34.82 ± 5.8
	Asia	30.4 ± 7.1	35.1 ± 7.2	33.4 ± 7.0	33.3 ± 7.3	33.0 ± 6.9	36.4 ± 7.1	35.9 ± 8.5
	Europe	29.7 ± 7.6	36.1 ± 8.0	32.4 ± 7.7	34.5 ± 7.8	31.9 ± 6.8	35.5 ± 8.3	37.6 ± 8.5
	North America	28.0 ± 7.2	36.7 ± 8.4	34.2 ± 5.6	35.6 ± 10.2	32.9 ± 8.3	35.6 ± 9.8	38.6 ± 10.3
	Oceania	30.9 ± 6.1	33.8 ± 8.1	30.7 ± 6.2	34.8 ± 7.9	31.4 ± 7.0	38.0 ± 8.0	36.8 ± 8.9
Family	Africa	30.9 ± 4.6	31.2 ± 8.8	33.5 ± 8.0	29.8 ± 8.5	31.9 ± 7.6	39.6 ± 9.2	38.3 ± 6.7
	Asia	28.8 ± 6.8	34.0 ± 8.8	31.6 ± 8.2	34.2 ± 8.4	30.8 ± 8.1	36.9 ± 9.4	35.6 ± 9.5
	Europe	30.9 ± 6.8	37.6 ± 7.4	33.8 ± 7.4	35.5 ± 8.0	39.3 ± 8.3	39.3 ± 8.3	40.6 ± 8.3
	North America	28.0 ± 5.4	35.2 ± 7.7	31.6 ± 6.9	32.1 ± 7.6	29.9 ± 5.0	36.8 ± 7.3	37.9 ± 6.6
	Oceania	33.3 ± 4.8	41.5 ± 6.0	33.5 ± 7.4	35.2 ± 8.8	28.9 ± 7.5	40.9 ± 6.4	44.0 ± 8.5
	South America	27.7 ± 6.5	32.2 ± 9.1	31.6 ± 7.8	31.2 ± 9.0	29.3 ± 6.2	37.7 ± 8.8	35.8 ± 8.7
Health	Africa	29.9 ± 6.4	34.4 ± 5.8	31.8 ± 7.4	28.1 ± 4.4	27.5 ± 7.0	30.2 ± 8.0	30.2 ± 6.4
	Asia	31.7 ± 6.6	35.9 ± 8.7	32.9 ± 8.1	32.3 ± 7.7	32.4 ± 7.0	36.3 ± 8.7	35.1 ± 9.1
	Europe	32.5 ± 7.0	36.6 ± 8.0	34.8 ± 7.4	34.8 ± 7.4	32.5 ± 7.7	36.7 ± 9.0	36.2 ± 9.4
	North America	27.8 ± 9.3	30.4 ± 8.9	31.3 ± 7.8	28.3 ± 7.9	30.7 ± 8.2	32.6 ± 7.5	31.0 ± 8.5
	Oceania	31.2 ± 7.8	36.4 ± 8.2	34.5 ± 8.7	35.2 ± 8.0	33.1 ± 8.7	37.4 ± 8.4	35.8 ± 7.8
National Identity	Asia	34.3 ± 7.6	32.3 ± 7.8	31.2 ± 8.5	29.9 ± 8.8	33.8 ± 7.8	37.9 ± 9.0	35.2 ± 8.8
	Europe	33.6 ± 8.2	34.8 ± 8.1	33.7 ± 7.6	33.5 ± 7.5	33.4 ± 8.2	38.3 ± 8.1	36.7 ± 8.2
	North America	32.4 ± 9.9	31.7 ± 6.3	32.2 ± 7.3	30.9 ± 7.6	32.0 ± 8.0	38.3 ± 9.4	35.3 ± 9.1
Religion	Africa	27.1 ± 9.7	31.5 ± 11.7	28.0 ± 10.4	29.3 ± 10.3	27.2 ± 8.2	31.7 ± 9.0	34.4 ± 10.6
	Asia	34.4 ± 7.9	35.9 ± 7.9	36.7 ± 6.7	36.6 ± 8.5	35.1 ± 7.2	39.7 ± 9.0	38.5 ± 9.2
	Europe	38.4 ± 8.2	42.0 ± 8.8	38.7 ± 8.7	40.6 ± 8.2	38.8 ± 8.1	42.7 ± 9.3	42.5 ± 8.6
	North America	31.4 ± 9.1	37.7 ± 8.3	36.3 ± 8.9	36.7 ± 9.9	32.5 ± 9.5	37.8 ± 9.1	38.3 ± 11.5
	South America	31.4 ± 7.7	34.6 ± 9.8	33.9 ± 7.9	34.6 ± 8.5	31.0 ± 7.8	37.9 ± 8.1	34.3 ± 9.9
Role of Government	Africa	33.4 ± 8.6	34.8 ± 4.5	33.6 ± 6.1	32.8 ± 6.7	32.9 ± 5.7	36.1 ± 4.9	36.5 ± 6.2
	Asia	32.8 ± 7.3	34.1 ± 7.0	32.6 ± 8.0	33.1 ± 8.3	33.1 ± 6.8	36.5 ± 7.2	36.8 ± 7.3
	Europe	35.3 ± 6.7	37.6 ± 7.3	35.9 ± 8.3	36.9 ± 8.3	35.9 ± 7.3	40.6 ± 8.1	41.1 ± 8.0
	North America	36.7 ± 6.1	35.4 ± 4.7	34.2 ± 6.0	38.0 ± 5.5	35.9 ± 7.6	37.8 ± 8.9	38.8 ± 6.8
	Oceania	38.2 ± 8.0	37.3 ± 8.7	35.5 ± 5.3	35.6 ± 6.9	33.7 ± 6.2	39.7 ± 7.1	42.4 ± 6.2
	South America	34.4 ± 7.2	35.5 ± 6.8	31.1 ± 7.9	32.1 ± 7.7	36.9 ± 6.8	37.3 ± 6.4	39.0 ± 8.2
Social Inequality	Africa	23.6 ± 8.4	28.5 ± 9.1	26.3 ± 8.0	25.4 ± 6.5	27.4 ± 6.9	26.4 ± 6.8	29.6 ± 9.6
	Asia	30.2 ± 9.0	34.2 ± 8.2	29.3 ± 8.8	31.7 ± 7.8	30.6 ± 8.4	34.0 ± 8.2	36.0 ± 9.2
	Europe	31.0 ± 8.2	36.4 ± 9.2	31.8 ± 9.0	34.6 ± 9.0	32.2 ± 8.6	37.1 ± 9.1	37.8 ± 10.2
	North America	31.3 ± 7.3	35.4 ± 7.9	31.6 ± 9.5	32.2 ± 8.7	29.7 ± 8.1	36.0 ± 9.2	37.6 ± 9.5
	Oceania	33.2 ± 7.4	37.9 ± 6.7	34.4 ± 6.7	35.5 ± 8.0	34.1 ± 8.0	37.0 ± 8.9	39.6 ± 8.5
	South America	27.9 ± 7.6	31.2 ± 8.2	27.4 ± 8.3	29.7 ± 8.5	29.3 ± 7.9	34.0 ± 9.0	31.8 ± 9.6
Social Networks	Africa	33.9 ± 7.2	39.6 ± 7.1	35.6 ± 7.8	39.3 ± 9.7	32.3 ± 7.9	35.9 ± 9.6	37.4 ± 8.6
	Asia	34.0 ± 7.3	38.3 ± 9.1	35.4 ± 9.7	36.4 ± 10.3	35.0 ± 8.2	35.7 ± 8.5	37.2 ± 8.8
	Europe	33.3 ± 7.3	37.9 ± 8.6	33.9 ± 8.5	34.7 ± 8.8	35.1 ± 7.5	36.6 ± 7.3	36.6 ± 7.4
	North America	32.0 ± 6.3	38.5 ± 8.0	35.2 ± 6.9	36.7 ± 7.3	35.8 ± 7.6	35.9 ± 7.0	36.3 ± 7.8
	Oceania	33.1 ± 6.8	38.4 ± 6.7	34.6 ± 8.0	34.6 ± 7.5	35.0 ± 7.0	36.6 ± 6.0	37.4 ± 6.8
	South America	31.6 ± 7.4	35.9 ± 8.2	31.8 ± 7.8	31.2 ± 9.0	33.7 ± 7.0	31.4 ± 6.2	33.4 ± 6.1
Work Orientations	Africa	31.3 ± 9.0	36.0 ± 8.0	33.2 ± 6.0	35.7 ± 5.9	31.5 ± 9.0	38.8 ± 6.6	
	Asia	30.9 ± 7.8	36.3 ± 7.6	33.6 ± 7.6	34.6 ± 7.6	32.4 ± 6.9	37.0 ± 7.7	38.0 ± 9.1
	Europe	32.7 ± 6.6	38.1 ± 7.4	34.7 ± 7.1	35.7 ± 7.6	33.6 ± 6.2	38.8 ± 7.6	39.5 ± 7.8
	North America	33.7 ± 5.7	35.2 ± 6.5	33.7 ± 6.4	33.5 ± 7.9	35.4 ± 5.9	42.4 ± 7.9	36.8 ± 7.4
	Oceania	32.6 ± 6.7	39.0 ± 7.0	34.2 ± 7.8	34.4 ± 7.0	34.3 ± 7.6	40.5 ± 8.2	42.0 ± 6.8
	South America	29.4 ± 6.3	34.4 ± 7.0	32.1 ± 9.1	34.2 ± 8.1	32.4 ± 7.6	38.5 ± 7.0	35.1 ± 7.2

Table 14: Mean accuracy (% $\pm$ SD) of each model across gender groups for all domains.

Domain	Gender	Qwen2.5-7B	Qwen2.5-72B	Qwen3-8B	Qwen3-32B	Llama-3.1-8B	Llama-3.3-70B	GPT-4o
Citizenship	Female	41.9 ± 9.0	44.3 ± 8.2	40.7 ± 8.3	44.6 ± 7.9	41.2 ± 8.1	44.2 ± 8.5	44.4 ± 9.2
	Male	39.9 ± 8.3	42.9 ± 7.6	39.9 ± 7.8	42.6 ± 7.5	40.0 ± 7.5	43.8 ± 9.0	44.2 ± 8.8
Environment	Female	30.9 ± 7.0	36.2 ± 7.4	33.1 ± 7.7	34.4 ± 7.9	32.8 ± 7.0	36.6 ± 7.7	38.1 ± 7.5
	Male	28.8 ± 7.5	34.8 ± 8.4	32.3 ± 7.0	33.8 ± 7.7	31.3 ± 6.5	35.3 ± 8.3	36.0 ± 9.3
Family	Female	30.9 ± 6.9	37.5 ± 8.5	33.7 ± 7.6	34.9 ± 8.4	33.0 ± 6.9	38.6 ± 8.7	39.9 ± 8.8
	Male	29.3 ± 6.4	35.0 ± 7.5	32.5 ± 7.5	34.2 ± 8.1	30.8 ± 7.0	38.6 ± 8.3	38.3 ± 8.4
Health	Female	31.5 ± 6.8	35.9 ± 8.4	34.4 ± 7.6	33.7 ± 7.6	31.5 ± 7.5	36.0 ± 8.5	35.1 ± 9.2
	Male	32.2 ± 7.6	35.9 ± 8.2	33.5 ± 7.9	33.5 ± 7.9	33.0 ± 7.8	36.4 ± 9.2	35.6 ± 9.3
National Identity	Female	33.3 ± 7.7	34.2 ± 8.5	33.2 ± 7.7	32.3 ± 7.9	33.4 ± 8.1	37.7 ± 8.2	35.8 ± 8.4
	Male	34.0 ± 8.5	34.1 ± 7.6	33.0 ± 7.9	33.0 ± 8.0	33.4 ± 8.2	38.6 ± 8.4	36.8 ± 8.4
Religion	Female	36.6 ± 8.5	40.0 ± 9.4	38.2 ± 8.2	38.8 ± 8.2	37.4 ± 8.2	41.3 ± 9.2	41.0 ± 8.8
	Male	36.4 ± 8.9	39.6 ± 9.3	36.9 ± 9.0	39.0 ± 9.4	36.5 ± 8.8	41.2 ± 9.7	40.5 ± 9.8
Role of Government	Female	35.0 ± 7.1	36.8 ± 6.7	35.1 ± 8.2	35.8 ± 8.0	35.7 ± 7.4	39.3 ± 7.4	40.1 ± 8.0
	Male	34.7 ± 7.0	36.3 ± 7.7	34.2 ± 8.1	35.3 ± 8.5	34.9 ± 7.0	39.1 ± 8.3	39.6 ± 8.0
Social Inequality	Female	30.3 ± 8.7	35.0 ± 9.2	30.6 ± 9.4	33.1 ± 9.0	31.6 ± 8.5	35.6 ± 8.9	35.9 ± 10.1
	Male	30.5 ± 8.0	35.3 ± 8.7	31.1 ± 8.3	33.3 ± 8.8	31.3 ± 8.3	35.8 ± 9.3	37.3 ± 10.0
Social Networks	Female	33.5 ± 7.4	38.7 ± 8.7	34.4 ± 8.6	35.6 ± 8.8	35.0 ± 7.6	36.5 ± 7.1	37.1 ± 7.8
	Male	33.2 ± 7.0	37.4 ± 8.2	34.4 ± 8.7	34.9 ± 9.4	35.0 ± 7.7	35.8 ± 8.1	36.2 ± 7.7
Work Orientations	Female	32.4 ± 6.9	37.4 ± 7.3	34.2 ± 7.2	35.5 ± 8.0	33.7 ± 6.5	38.8 ± 7.6	38.7 ± 7.9
	Male	31.9 ± 7.0	37.4 ± 7.5	34.2 ± 7.4	35.0 ± 7.1	33.1 ± 6.6	38.8 ± 7.6	39.1 ± 8.1

Table 15: Mean accuracy (% ± SD) of each model across age ranges for all domains.

Domain	Age Range	Qwen2.5-7B	Qwen2.5-72B	Qwen3-8B	Qwen3-32B	Llama-3.1-8B	Llama-3.3-70B	GPT-4o
Citizenship	18–25	41.5 ± 8.4	45.2 ± 7.1	40.5 ± 7.9	43.6 ± 8.6	39.8 ± 8.2	45.1 ± 8.6	44.0 ± 9.4
	26–35	40.2 ± 9.7	42.8 ± 7.4	39.9 ± 8.2	44.4 ± 8.0	39.8 ± 7.7	42.6 ± 8.4	43.3 ± 8.9
	36–45	40.7 ± 10.1	43.4 ± 8.9	40.2 ± 8.5	43.9 ± 7.7	39.8 ± 8.4	44.4 ± 8.7	43.7 ± 8.8
	46–55	39.3 ± 7.5	42.8 ± 7.6	40.7 ± 8.2	43.0 ± 7.2	40.5 ± 7.3	43.8 ± 8.5	43.6 ± 8.1
	56–65	42.0 ± 8.4	43.6 ± 7.8	39.6 ± 6.8	42.5 ± 8.4	40.9 ± 7.7	43.7 ± 9.4	45.2 ± 9.7
	66+	42.5 ± 7.9	44.7 ± 8.2	40.9 ± 8.3	44.5 ± 7.5	41.4 ± 8.0	44.8 ± 9.0	46.0 ± 9.3
Environment	18–25	32.1 ± 6.4	33.9 ± 6.8	31.8 ± 5.9	32.7 ± 7.1	33.0 ± 6.8	36.7 ± 7.0	38.0 ± 8.2
	26–35	30.2 ± 7.3	35.1 ± 8.6	32.2 ± 8.0	33.5 ± 8.8	31.9 ± 6.9	34.8 ± 8.4	36.9 ± 10.5
	36–45	28.2 ± 7.1	35.1 ± 7.3	33.9 ± 7.2	34.8 ± 6.9	31.5 ± 7.6	35.1 ± 7.5	36.9 ± 7.7
	46–55	30.0 ± 6.8	35.5 ± 7.6	32.4 ± 7.4	34.0 ± 8.1	32.2 ± 6.3	36.3 ± 8.0	36.4 ± 8.6
	56–65	30.0 ± 7.3	36.3 ± 8.0	32.8 ± 7.5	33.6 ± 7.4	32.3 ± 6.9	36.1 ± 8.8	37.3 ± 7.8
	66+	29.7 ± 8.2	36.1 ± 8.1	32.8 ± 7.5	35.7 ± 8.1	31.9 ± 6.7	36.9 ± 7.2	37.6 ± 8.0
Family	18–25	29.6 ± 3.9	35.1 ± 7.4	30.7 ± 8.7	31.3 ± 11.1	31.9 ± 6.6	39.6 ± 9.2	38.6 ± 10.1
	26–35	29.9 ± 6.9	34.3 ± 8.6	31.4 ± 6.4	32.2 ± 8.5	31.2 ± 7.1	35.8 ± 7.3	37.0 ± 8.3
	36–45	31.4 ± 6.5	36.3 ± 8.4	33.5 ± 8.2	34.4 ± 8.9	32.1 ± 7.2	38.3 ± 9.2	38.5 ± 9.0
	46–55	29.6 ± 7.0	35.6 ± 7.7	33.7 ± 8.1	33.9 ± 8.2	31.9 ± 7.0	38.6 ± 8.1	38.8 ± 8.8
	56–65	29.1 ± 6.8	36.7 ± 7.6	32.6 ± 7.2	35.5 ± 7.7	31.5 ± 6.5	38.2 ± 7.8	40.2 ± 8.6
	66+	30.6 ± 6.4	38.2 ± 8.5	33.5 ± 7.0	36.2 ± 7.2	32.6 ± 7.6	41.6 ± 9.0	40.6 ± 8.0
Health	18–25	31.9 ± 8.1	36.2 ± 7.8	32.9 ± 7.2	34.0 ± 8.1	31.3 ± 7.0	38.5 ± 8.2	34.9 ± 9.2
	26–35	30.9 ± 7.3	35.9 ± 7.7	33.0 ± 8.4	33.3 ± 7.9	31.8 ± 7.4	37.4 ± 8.7	34.3 ± 9.5
	36–45	30.8 ± 7.2	35.2 ± 8.1	34.0 ± 7.4	33.8 ± 7.7	31.4 ± 7.7	36.4 ± 8.8	36.0 ± 9.5
	46–55	27.8 ± 9.3	32.5 ± 8.8	32.0 ± 8.0	31.1 ± 7.4	29.5 ± 8.2	34.3 ± 9.0	34.9 ± 9.4
	56–65	31.2 ± 7.8	36.4 ± 8.2	34.5 ± 8.7	35.2 ± 8.0	32.5 ± 7.9	37.5 ± 9.3	34.5 ± 8.7
	66+	32.9 ± 8.4	37.3 ± 8.5	35.4 ± 9.0	35.5 ± 8.7	34.6 ± 8.5	38.6 ± 9.5	37.1 ± 8.9
National Identity	18–25	33.3 ± 7.7	34.7 ± 8.1	32.4 ± 7.9	31.7 ± 8.2	33.7 ± 8.0	38.5 ± 7.8	33.3 ± 8.3
	26–35	34.2 ± 7.8	34.9 ± 8.0	32.7 ± 8.3	32.5 ± 8.0	33.7 ± 8.1	39.2 ± 8.3	35.9 ± 8.5
	36–45	33.7 ± 8.3	35.5 ± 7.9	33.3 ± 7.5	32.8 ± 7.4	33.9 ± 7.6	39.4 ± 8.4	38.0 ± 7.3
	46–55	32.4 ± 9.9	34.8 ± 8.4	32.4 ± 7.7	32.1 ± 7.8	33.3 ± 7.8	38.8 ± 9.2	36.9 ± 9.0
	56–65	34.0 ± 7.8	35.6 ± 8.2	33.6 ± 7.9	33.1 ± 8.0	33.8 ± 7.9	38.2 ± 8.5	36.0 ± 8.3
	66+	34.4 ± 8.4	35.9 ± 8.0	33.3 ± 8.6	32.8 ± 8.6	34.5 ± 8.1	39.5 ± 9.0	37.2 ± 8.2
Religion	18–25	34.7 ± 9.7	38.9 ± 10.7	35.4 ± 8.4	37.5 ± 10.1	35.5 ± 8.5	39.7 ± 9.9	39.6 ± 10.4
	26–35	35.8 ± 7.9	38.0 ± 9.2	37.3 ± 9.4	37.9 ± 8.9	35.2 ± 7.9	41.8 ± 9.3	39.4 ± 9.2
	36–45	36.9 ± 8.0	40.3 ± 8.7	37.5 ± 8.2	39.2 ± 8.3	38.1 ± 7.9	41.5 ± 8.8	41.7 ± 9.5
	46–55	36.6 ± 8.2	40.2 ± 8.8	37.7 ± 8.6	38.1 ± 8.4	37.7 ± 9.3	41.3 ± 9.3	41.4 ± 8.0
	56–65	38.2 ± 9.9	40.9 ± 10.1	38.8 ± 8.9	40.5 ± 9.2	38.2 ± 8.7	42.6 ± 9.4	42.1 ± 10.2
	66+	35.9 ± 8.3	39.6 ± 8.7	38.0 ± 8.0	39.6 ± 8.1	36.0 ± 8.0	39.8 ± 10.2	39.1 ± 8.3
Role of Government	18–25	34.7 ± 6.5	37.0 ± 6.2	35.6 ± 8.7	37.1 ± 8.1	36.8 ± 7.5	40.6 ± 7.6	41.2 ± 7.0
	26–35	34.1 ± 7.2	36.1 ± 7.6	33.5 ± 8.5	34.6 ± 9.7	34.8 ± 6.9	38.4 ± 8.2	39.4 ± 8.1
	36–45	33.9 ± 5.7	35.6 ± 7.0	32.5 ± 8.1	34.2 ± 7.8	36.1 ± 6.5	38.4 ± 6.9	39.0 ± 7.8
	46–55	35.4 ± 7.7	36.5 ± 8.4	35.3 ± 8.0	36.3 ± 7.8	35.7 ± 7.3	39.1 ± 7.2	39.5 ± 8.5
	56–65	36.9 ± 6.3	36.9 ± 6.0	36.6 ± 7.5	36.3 ± 8.3	34.8 ± 6.5	39.6 ± 7.3	41.6 ± 8.1
	66+	33.9 ± 7.9	37.1 ± 7.4	35.1 ± 7.7	34.5 ± 7.2	33.8 ± 8.0	39.0 ± 9.4	38.4 ± 7.2
Social Inequality	18–25	28.5 ± 8.3	34.8 ± 7.6	30.6 ± 6.8	33.6 ± 7.8	30.3 ± 7.8	34.9 ± 6.7	34.9 ± 11.1
	26–35	30.6 ± 7.6	35.2 ± 9.3	30.8 ± 7.6	33.7 ± 7.5	31.8 ± 7.8	33.6 ± 8.3	35.5 ± 8.0
	36–45	30.6 ± 8.4	35.2 ± 9.0	31.2 ± 8.7	33.2 ± 8.6	31.3 ± 8.9	36.8 ± 9.3	38.0 ± 10.4
	46–55	30.2 ± 8.8	35.3 ± 9.0	31.3 ± 8.5	33.1 ± 8.3	31.7 ± 7.8	36.7 ± 9.6	37.1 ± 10.5
	56–65	31.5 ± 8.7	35.1 ± 9.3	30.9 ± 9.8	33.2 ± 9.6	31.7 ± 9.2	34.9 ± 9.6	36.1 ± 10.6
	66+	30.0 ± 8.0	35.2 ± 9.0	30.5 ± 10.2	32.8 ± 10.6	31.4 ± 8.8	36.6 ± 9.2	36.9 ± 9.8
Social Networks	18–25	33.0 ± 7.3	37.1 ± 8.0	32.4 ± 8.8	34.3 ± 10.4	34.3 ± 7.9	36.6 ± 7.1	38.2 ± 7.0
	26–35	33.9 ± 7.2	39.1 ± 8.3	34.8 ± 8.0	36.2 ± 8.3	35.4 ± 8.2	36.6 ± 8.7	37.6 ± 8.5
	36–45	32.8 ± 6.2	38.9 ± 8.0	33.4 ± 8.5	34.2 ± 7.9	34.8 ± 7.7	37.0 ± 6.6	36.9 ± 7.0
	46–55	32.4 ± 7.0	37.2 ± 8.6	34.4 ± 8.8	34.9 ± 9.9	35.7 ± 7.7	35.8 ± 8.6	36.4 ± 8.2
	56–65	32.6 ± 7.2	36.6 ± 8.9	34.6 ± 8.1	34.7 ± 8.8	33.9 ± 7.0	34.4 ± 7.3	34.4 ± 7.3
	66+	36.6 ± 7.9	40.5 ± 8.3	37.4 ± 8.9	39.2 ± 8.3	36.6 ± 7.2	37.1 ± 7.3	37.9 ± 8.2
Work Orientations	18–25	32.9 ± 7.0	35.9 ± 6.5	33.5 ± 6.5	33.9 ± 8.1	32.4 ± 7.4	38.0 ± 7.0	37.5 ± 6.9
	26–35	31.2 ± 6.6	35.9 ± 6.8	33.7 ± 7.6	35.0 ± 8.3	33.1 ± 6.6	38.0 ± 7.4	38.7 ± 8.2
	36–45	32.1 ± 7.2	37.9 ± 7.5	33.8 ± 7.1	35.2 ± 7.3	33.4 ± 6.4	38.6 ± 7.7	39.6 ± 8.5
	46–55	33.2 ± 6.9	38.4 ± 8.0	35.0 ± 8.0	36.0 ± 7.7	34.6 ± 6.3	40.3 ± 7.9	39.4 ± 8.0
	56–65	31.8 ± 6.8	37.2 ± 7.2	34.0 ± 6.4	34.9 ± 7.0	32.3 ± 7.0	38.1 ± 7.2	38.0 ± 7.3
	66+	32.7 ± 6.8	40.1 ± 6.3	39.5 ± 6.6	37.5 ± 4.9	32.5 ± 4.3	39.7 ± 9.4	41.4 ± 5.0