SRP: Understanding Reliability in LLM-based Human Behavior Simulation

Anonymous ACL submission

Abstract

Large language models (LLMs) are increasingly used to simulate human survey responses and behavioral reactions, yet the conditions under which such simulations are reliable remain unclear, making it difficult to pinpoint where errors arise and which configuration choices drive them. To make reliability analysis more interpretable and actionable, we propose the Simulation Reliability Prism (SRP), which decomposes simulation into three structured layers and analyzes error propagation across layers along three key configuration dimensions—model capacity, profile completeness, and population coverage, while jointly evaluating two complementary targets: individual-level reliability and population-level reliability. Across three survey tasks and eleven LLMs, we show that profile conditioning is necessary to avoid systematic distributional bias, while increasing profile completeness yields diminishing individual gains and transfers unreliably to population-level improvements, sometimes reversing. Increasing population coverage mainly reduces variance, and population-level reliability typically stabilizes with fewer than 100 samples. Our findings offer practical guidance for reliable LLM-based survey simulation.[1]

1 Introduction

Human experiments and surveys are fundamental tools for studying human behavior, but they are costly and time-consuming(Aher et al., 2023; Argyle et al., 2023; Bisbee et al., 2024). Recent advances in large language models (LLMs) have sparked interest in using them as computational proxies for human participants, offering rapid, low-cost behavioral simulation at scale (Dominguez-Olmedo et al., 2024; Hu et al., 2024; Manning et al., 2024; Binz et al., 2025a; Hu et al., 2025).

Early explorations suggest promise: LLMs can reproduce certain patterns of human decision-making across economic games, survey responses, and social judgments (Aher et al., 2023; Ahnert et al., 2025; Hewitt et al., 2024). However, a fundamental question remains: **when can we trust LLM-based simulations to faithfully represent human behavior?** Despite growing adoption, there is limited systematic understanding of what determines simulation reliability. Existing work typically evaluates isolated components—such as prompting strategies (Cho et al., 2024; Hewitt et al., 2024), persona engineering (Cho et al., 2024; Moon et al., 2024), or alignment techniques (Chu et al., 2023; Suh et al., 2025; Binz et al., 2025b; Kolluri et al., 2025)—within narrow settings, leaving practitioners without clear principles for when simulations will succeed or fail.

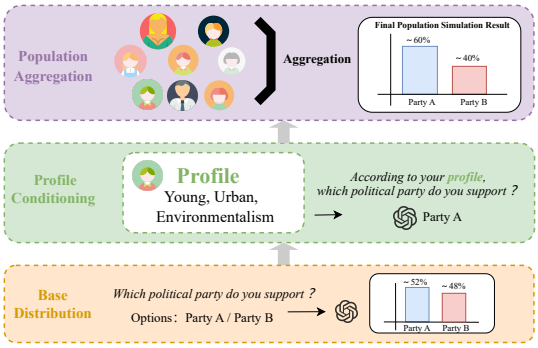


Figure 1: LLM-based Human Behavior Simulation As a Hierarchical Structure.

Our key insight is that human behavior simulation emerges from a hierarchical structure where each layer builds upon the previous one. As illustrated in Figure 1, taking party support prediction in a survey as an example, the structure unfolds in three layers: (1) **Base distribution**: The model induces an intrinsic behavioral distribution shaped by pretraining and alignment, capturing its default tendency before seeing any individual information. (2) **Profile conditioning**: The base distribution is conditioned on profile information such as demographics, values, or prior experiences, producing dif-

[1]We will release our data and code after blind review.

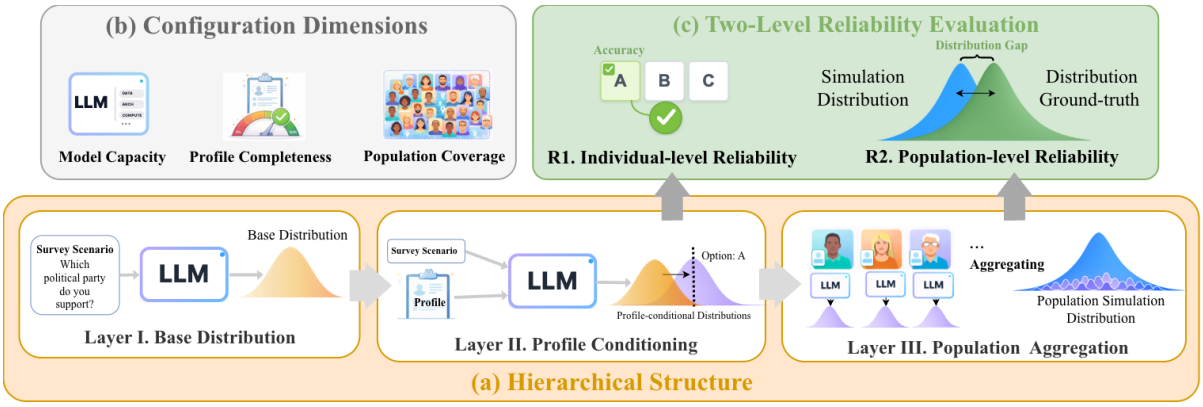


Figure 2: Simulation Reliability Prism Overview: Three-Layer Structure with Configurable Dimensions and Two-Level Reliability Evaluation.

ferentiated behavioral tendencies across personas. (3) **Population aggregation**: Individual responses compose into collective patterns, where sampling strategy and coverage determine how diverse perspectives combine. **These three layers jointly determine simulation reliability, yet their relative contributions and interactions remain poorly understood.** Yet most existing evaluations only observe final accuracy without decomposing where errors originate, how layers interact, or which components most constrain performance.

To address this gap, we propose the **S**imulation **R**eliability **P**rism (SRP), a systematic framework that decomposes simulation into three structured layers and analyzes reliability across three configuration dimensions—**model capacity**,**profile completeness**, and **population coverage**, while evaluating reliability at two complementary levels: **individual-level**(R1) and **population-level**(R2). SRP turns reliability from a single score into a structured analysis of underlying mechanisms and error propagation across layers and dimensions.

Guided by the SRP, we conduct experiments on three survey tasks with eleven LLMs. Our analysis reveals four critical patterns: (1) Without profile conditioning, models exhibit substantial population-level distributional bias that varies widely across models and tasks. (2) Increasing profile completeness improves R1 with diminishing returns; larger and more capable models benefit more from richer profiles. (3) At low-to-moderate profile completeness, R1 and R2 generally improve together; at high completeness, this coupling weakens and R1 gains no longer reliably translate to R2 improvements. (4) Increasing population coverage primarily reduces variance rather than systematic bias, and R2 typically stabilizes when coverage reaches about 50–100 samples. Collectively, these findings motivate three priorities for future work: **increasing the informativeness of profile attributes; strengthening models capacity to integrate and leverage complex profiles; and explicitly optimizing population-level distributional alignment as an objective.**

Our contributions are:

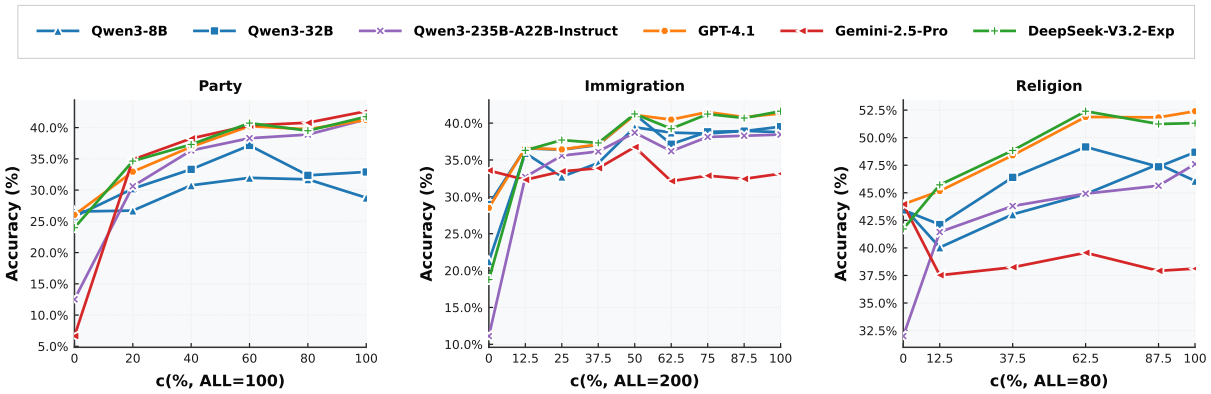
(1) We introduce the SRP, which transforms reliability from a single score into a decomposable analysis across three layers, three configuration dimensions, and two evaluation levels.

(2) We conduct systematic experiments across three survey tasks and eleven LLMs, examining how these configuration dimensions shape reliability at each simulation layer.

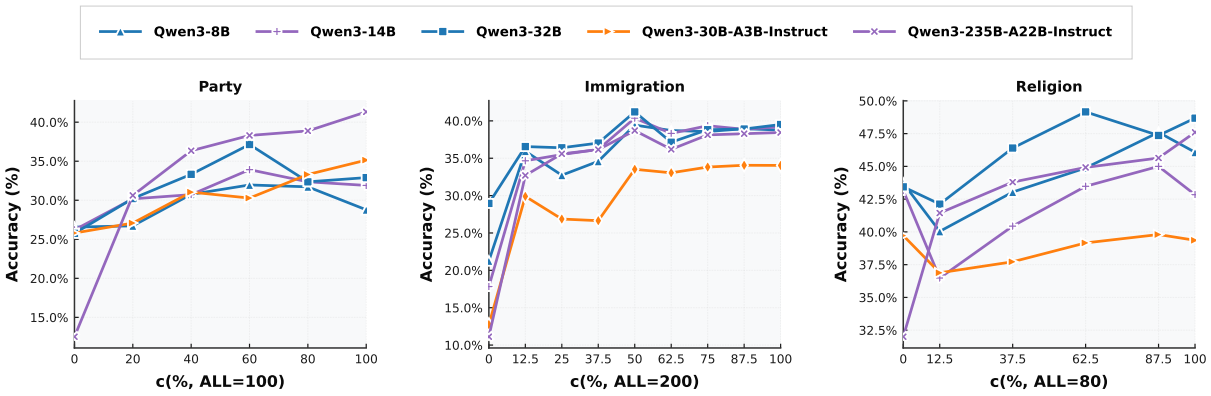
(3) Building on the empirical evidence, we discuss key challenges in current LLM-based human behavior simulation and outline actionable directions for future research.

Model	Party	Immigration	Religion
GPT-4.1 / Gemini			
GPT-4.1	0.670	0.368	0.542
Gemini-2.5-Pro	0.589	0.297	0.549
DeepSeek			
DeepSeek-V3.2-Exp	0.483	**0.128**	0.342
DeepSeek-R1	0.664	0.579	0.345
Qwen3 series			
Qwen3-8B	0.689	0.304	0.540
Qwen3-14B	0.722	0.326	0.500
Qwen3-32B	0.664	0.408	0.474
Qwen3-30B-A3B-Instruct	0.378	0.396	0.403
Qwen3-30B-A3B-Thinking	0.527	0.181	0.443
Qwen3-235B-A22B-Instruct	**0.271**	0.482	0.263
Qwen3-235B-A22B-Thinking	0.475	0.578	**0.173**

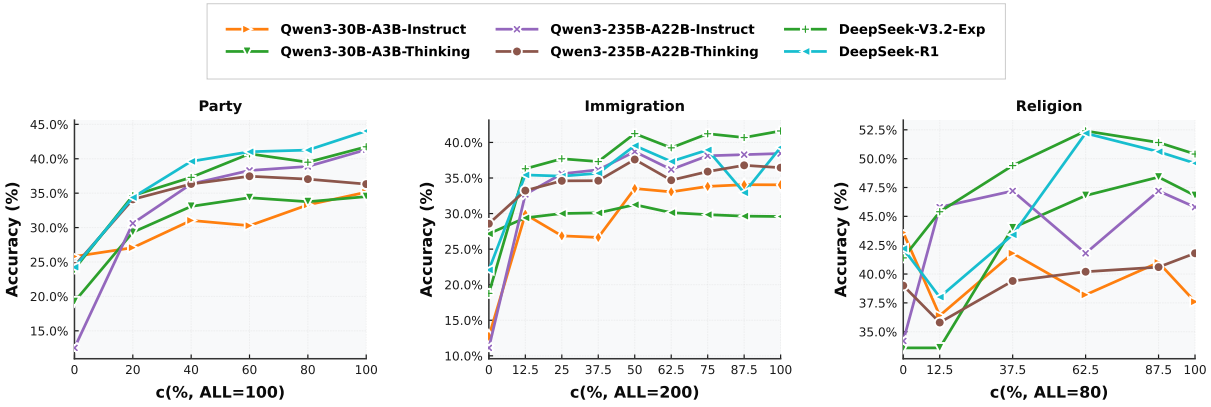
Table 1: TVD↓ between LLM-induced base response distributions and human survey distributions.



(a) c-Acc curves across model families



(b) c-ACC curves across model scales



(c) c-ACC curves across reasoning variants

Figure 3: Changes in individual-level simulation accuracy as profile completeness c increases across models. The x-axis c denotes the profile attribute coverage rate, i.e., the percentage of attributes provided to the model out of all available attributes in the dataset.

2 Simulation Reliability Prism

2.1 Simulation Hierarchical Structure

We characterize LLM-based human behavior simulation as three structured layers, and define the distribution induced by each layer, as illustrated in Figure 2(a).

Layer I: Base distribution. The first layer captures the model's default responses without any personal profile information. It is the baseline distribution induced under a given scenario. This distribution is mainly shaped by the model architecture, pretraining data, and alignment strategy. Formally, let f_ψ denote a pretrained LLM with parameters ψ. Given a scenario s, we define the base distribution as:

$$\mathbb{P}_s \triangleq P_{f_\psi}(y \mid s). \tag{1}$$

Layer II: Profile conditioning. Building on the baseline distribution, the model is provided with an individual profile, such as geographic information, psychological tendencies, and past behaviors,

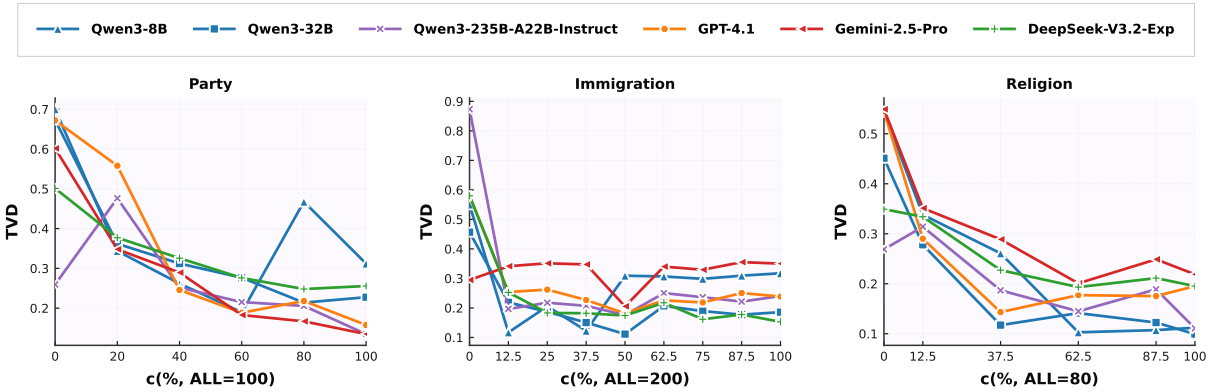


Figure 4: Effect of profile completeness on population simulation reliability.

yielding a profile-conditioned distribution. Formally, given an individual profile x, we define the profile-conditional outcome distribution as

$$\mathbb{P}_{s,x} \triangleq P_{f_\psi}(y \mid s, x). \tag{2}$$

Layer III: Population aggregation. To obtain population-level simulation results, the third layer aggregates the simulated outputs of sampled individuals to form a population distribution. Formally, we consider a finite population panel of N individuals $\mathcal{X}_N = \{x_i\}_{i=1}^N$. Aggregating individualized distributions of $\mathcal{X}_N$ yields the population-aggregated outcome distribution:

$$\mathbb{P}_{s,\mathcal{X}_N} \triangleq \frac{1}{N}\sum_{i=1}^{N} P_{f_\psi}(y \mid s, x_i). \tag{3}$$

Through the sequential generation and propagation from Layer I to Layer III, a standard LLM-based human behavior simulation can be completed.

2.2 Configuration Dimensions

Building on the three-layer simulation structure, we further define three configuration dimensions to systematically study key variables in the simulation setup and analyze how they influence simulation reliability across layers. As shown in Figure 2(b), three configuration dimensions are:

Model capacity ($\mathcal{M}$). It concerns the choice of the base model, such as the model architecture and model size. This choice affects the simulation from Layer I and continues to propagate to later layers.

Profile completeness (c). It captures how complete the available user information is. In principle, richer profiles should enable more accurate simulation of individual behavior. Yet in LLM-based simulation, the effect need not be linear. We therefore study how profile completeness shapes reliability, including when it matters and how strongly. It affects Layer II and propagates to Layer III.[2]

Population coverage (N). It refers to the coverage of the sample used in population simulation. Intuitively, broader coverage should lead to more reliable population-level results. However, this relationship need to be quantified. We ask how much coverage is sufficient for a reliable population simulation, and how much additional gain we get from increasing coverage further. It affects Layer III.

2.3 Two-Level Reliability Evaluation

We assess the reliability of LLM-based human behavior simulation at two levels: the individual level (R1) and the population level (R2) as shown in Figure 2(c). Intuitively, stronger R1 should translate into stronger R2. However, this relationship is not guaranteed, because the transition from Layer II to Layer III is shaped by both profile completeness and population coverage. Profile completeness often improves R1, yet its effect on R2 can be more subtle. Meanwhile, population coverage determines how representative the aggregated sample is; insufficient coverage may introduce systematic bias and distort R2. As a result, it remains unclear how these two configuration dimensions modulate the strength and even the direction of the association between R1 and R2.

2.4 Theoretical Insight

We aim to characterize when an LLM-based simulator can be reliable at the *population level*. Let s denote a task scenario, u an individual, x an in-

[2]We quantify c as the proportion of attributes provided in a given profile relative to the total number of available attributes.

Table 2: Incremental changes in ACC and TVD across c.

	GPT-4.1		DeepSeek-V3.2-Exp		Qwen3-235B-A22B-Instruct	
c	ΔACC(c) ↑	ΔTVD(c) ↓	ΔACC(c) ↑	ΔTVD(c) ↓	ΔACC(c) ↑	ΔTVD(c) ↓
20	24.52%	-55.52%	39.78%	-43.66%	144.05%	5.91%
40	14.92%	-13.78%	12.41%	-16.99%	18.77%	-23.62%
60	6.36%	-31.88%	6.68%	6.33%	2.63%	-39.73%
80	1.42%	14.64%	-0.26%	-1.33%	6.35%	42.54%
100	2.71%	-22.73%	4.54%	5.12%	4.92%	-18.26%

dividual profile, and y a behavioral outcome. The target behavior distribution under scenario s is

$$p^*(y \mid s) = \sum_u p^*(u) \sum_x p^*(x \mid u)\, p^*(y \mid s, x, u). \tag{4}$$

An LLM-based simulator induces an approximate distribution of the same form,

$$p(y \mid s) = \sum_u p(u) \sum_x p(x \mid u)\, p(y \mid s, x, u), \tag{5}$$

and we define the population-level discrepancy as the ℓ_1 distance

$$D(s) \triangleq \left\| p^*(\cdot \mid s) - p(\cdot \mid s) \right\|_1. \tag{6}$$

Key idea: population mismatch is multi-source. It makes clear that a population distribution can deviate from the target for: (i) the simulated population composition is shifted, (ii) the conditional profile distribution is mismatched, or (iii) even given the same (s, x, u), the simulator produces incorrect conditional behavior. Crucially, these failure modes are *not interchangeable*: reducing one does not automatically compensate for the others.

A sufficient condition and its interpretation. Through successive substitution and the triangle inequality, one can show that $D(s) \to 0$ is guaranteed when the following component distributions converge:

$$\begin{aligned} p(u) &\to p^*(u), \\ p(x \mid u) &\to p^*(x \mid u), \\ p(y \mid s, x, u) &\to p^*(y \mid s, x, u). \end{aligned} \tag{7}$$

Detailed derivation is shown in Appendix A.

3 Experiments

3.1 Setup

Data. To evaluate LLM-based human behavior simulation across behavioral domains, we conduct experiments on three tasks: *religious stance*, *partisan preference*, and *immigration attitude*. We use established survey datasets widely adopted in social science: the European Social Survey (ESS; *Party*)(ESS ERIC, 2025), the World Values Survey (WVS; *Immigration*)(Association, 2022) and SocioBench (Socio; *Religion*)(Wang et al., 2025b). Detailed dataset descriptions and dataset sizes are reported in the Appendix B.

Metric. For R1, we use accuracy (ACC) to compare model-predicted behaviors with the ground-truth behaviors in the data. For R2, we use total variation distance (TVD) to quantify the gap between the simulated population behavior distribution and the ground-truth label distribution.

Model. We evaluate a diverse set of LLMs to study how model capacity affects the reliability of LLM-based human behavior simulation. We include both proprietary and open-source models: the proprietary models include GPT-4.1 (OpenAI, 2025) and Gemini-2.5 Pro (Comanici et al., 2025); the open-source models include DeepSeek-V3.2-Exp (DeepSeek, 2025), DeepSeek-R1 (Guo et al., 2025), and multiple models from the Qwen3 family (Yang et al., 2025) covering different parameter scales and reasoning variants.

3.2 Layer I — Base distribution

In Layer I, we examine the model's base behavioral tendency without profile conditioning. Specifically, we provide no profile information and only input the task context, making one prediction for each instance in the full dataset (for a total of N predictions, where N is the number of samples). We then compare the LLM-induced *base response distribution* with the *ground-truth distribution* from the human survey. Table 1 shows the results.

TVD scores exhibit substantial variation across models and tasks. Even within a single task, models can differ markedly in distributional error. For example, on *Party*, the score of Qwen3-14B and Qwen3-235B-A22B-Instruct differs by approximately 0.45. Moreover, population-level distribution matching performance shows no consistent correlation with model scale or general capacity. This suggests that the ability to estimate population-level distributions constitutes a distinct capacity dimension. This mismatch likely arises from biases in pretraining corpora and the lack of explicit training or calibration of LLMs to align with empirical survey distributions.

3.3 Layer II — Profile Conditioning

The second layer introduces profiles to condition the model's output distribution. In this part, we

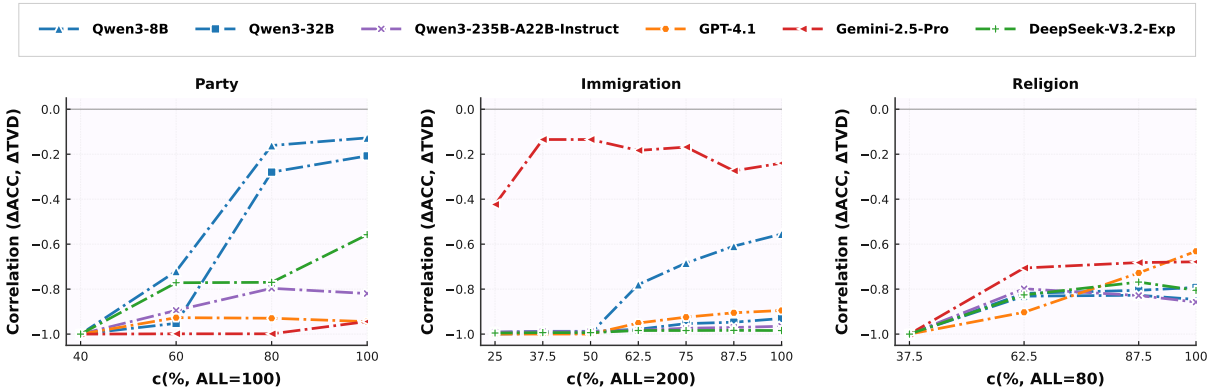


Figure 5: Relationship between individual-level accuracy and population-level reliability. We report the Pearson correlation between $\Delta\mathrm{ACC}(c') = \mathrm{ACC}(c') - \mathrm{ACC}(0)$ and $\Delta\mathrm{TVD}(c') = \mathrm{TVD}(c') - \mathrm{TVD}(0)$ across profile-completeness levels $c' \leq c$. Each point corresponds to an endpoint c'; more negative values indicate that higher ACC is associated with lower TVD.

focus on examining how model capacity and profile completeness influence this layer. Specifically, we progressively increase the completeness of the profile information and observe the changes in R1. The results are shown in Figure 3.

Across all tasks, we observe a consistent pattern: **Individual-level simulation accuracy increases with profile completeness, with diminishing marginal gains.** This aligns with our intuition that more complete user profiles provide models with richer individual-level information, enabling them to infer personal preferences that are closer to ground truth. On diminishing returns, it can be explained by two factors. On the one hand, as profile information accumulates, newly added attributes tend to be increasingly redundant. On the other hand, processing dense and high-dimensional profile context places greater demands on a model's contextual understanding and reasoning capacity, which further limits the effective utilization of additional information.

To further study how model capacity affects the diminishing returns of profile completeness, we compare different models.

Model Families. As shown in Figure 3a, we find that **performance trends at higher c are model-dependent**. As c grows large, some models' accuracy converges to a similar level and continues to increase such as Gemini-2.5-Pro, DeepSeek-V3.2-Exp, GPT-4.1 and Qwen3-235B-A22B-Instruct for *Party*, while others plateau at a lower level such as Qwen3-8B and Qwen3-32B, creating a clear performance split. GPT-4.1 and DeepSeek-V3.2-Exp consistently perform best across all three tasks; as c increases, their ACC continues to improve substantially, indicating strong potential for further gains.

Model Size. As shown in Figure 3b, we observe a broadly consistent trend across the three tasks: **at the same profile completeness c, larger models tend to achieve higher accuracy and continue to improve as c increases.** Within the Qwen3 family, Qwen3-32B consistently outperforms Qwen3-14B; under the MoE setting, Qwen3-235B-A22B-Instruct also generally outperforms Qwen3-30B-A3B-Instruct.

Instruct and Thinking variants. Figure 3c compares Instruct and Thinking models. We find no consistent evidence that Thinking models outperform Instruct models. Although Thinking models have been shown to demonstrate superior reasoning on mathematical and logical tasks, their reasoning abilities may differ from those needed for behavioral simulation. Simulating human behavior requires understanding social contexts, individual traits, and behavioral patterns that are potentially distinct from formal reasoning skills.

3.4 Layer III — Population Aggregation.

Layer III evaluates whether the aggregated population-level simulation outcomes can reproduce the true population distribution. We analyze: (i) how individual-level simulation accuracy from Layer II propagates to and affects population-level results, and (ii) how population coverage influences population-level simulation reliability.

Effect of Layer II Profile Completeness on Population-Level Reliability. We track R2 as Layer II profile completeness c increases in Figure 4. R2 improves with c but shows diminishing

returns, mirroring Layer II: as R1 increases, the aggregated distribution moves closer to the empirical one, and once R1 saturates, R2 largely plateaus.

As profile completeness increases, **accuracy gains do not always translate into population-level distributional alignment.** For small to moderate c, ΔACC and ΔTVD are strongly negatively correlated across tasks and models, indicating that accuracy improvements typically co-occur with TVD reductions. In the high-c regime, this correlation becomes less negative, suggesting a weakened coupling between individual- and population-level gains. Consistently, as shown in Table 2, we observe cases where accuracy improves but TVD worsens, e.g., GPT-4.1 at c=80 with $\Delta\text{ACC}= +1.42\%$ and $\Delta\text{TVD}= +14.64\%$, and Qwen3 at c=80 with $\Delta\text{ACC}= +6.35\%$ and $\Delta\text{TVD}= +42.54\%$. Figure 5 further quantifies this translation by reporting, for each endpoint c, the Pearson correlation between accuracy gains and TVD changes computed over completeness levels up to c.

Effect of Population Coverage on Population-Level Reliability. We study how population coverage affects R2 in Layer III by varying the number of simulated individuals N and measuring TVD between the aggregated simulated distribution and the target population distribution; see Figure 6.

Population reliability improves quickly as coverage increases and then levels off. Model rankings are also unstable at small N, but become much more consistent once N is large enough for TVD to stabilize. As N increases, the GT sampling baseline steadily decreases and approaches zero, whereas LLM-based simulations converge to a non-zero plateau. The persistent gap between the LLM curves and the GT baseline indicates that increasing N reduces sampling variance but cannot eliminate the residual distributional mismatch.

4 Discussion

Building on our experiments and analyses, we step back to revisit the most fundamental questions in LLM-based human behavior simulation: what drives reliability and where the key bottlenecks lie today. We present the following analysis and implications.

What drives reliability. We highlight that LLM-based human behavior simulation follows a hierarchical structure, moving from the model's base distribution to profile conditioning and then to population aggregation.

In *Layer I,* model capacity primarily determines simulation reliability. When only task context is provided without profile information, LLM-generated population distributions often deviate substantially from ground-truth surveys, with the magnitude of mismatch varying by model and task.

In *Layer II*, both model capacity and profile completeness jointly determine individual-level simulation performance. Increasing c generally improves accuracy, though with diminishing returns. Larger models show greater gains from richer profiles and reach higher performance ceilings.

In *Layer III*, population-level reliability depends on model capacity, profile completeness, and population coverage. Model capacity and profile completeness affect Layer III primarily by improving individual-level reliability in Layer II, which then propagates to aggregate outcomes. However, improvements in Layer II do not always translate to Layer III gains: as profile completeness increases, population-level metrics may plateau or even decline despite continued individual-level improvements. In contrast, population coverage mainly reduces sampling noise, with results stabilizing at relatively small sample sizes (typically around $N \approx 50$–100) and marginal improvements beyond this threshold.

Where the key bottlenecks lie today. *For individual-level simulation*, the current bottleneck lies not in the quantity of profile information, but in the informativeness of additional attributes and the model's capacity to integrate them effectively. Crucially, this integration capacity differs from the reasoning abilities that current models excel at—such as mathematical and logical reasoning—requiring instead social-cognitive understanding.

For population-level simulation, the key bottleneck is that higher individual accuracy does not reliably translate to better population distributions matching. Increasing sample size can only provide limited benefits. Together, these challenges indicate that population-level metrics should be directly optimized as independent targets.

Collectively, these findings motivate three priorities for future work: **increasing the informativeness of profile attributes**; **strengthening models capacity to integrate and leverage complex profiles**; and **explicitly optimizing population-level distributional alignment as an objective.**

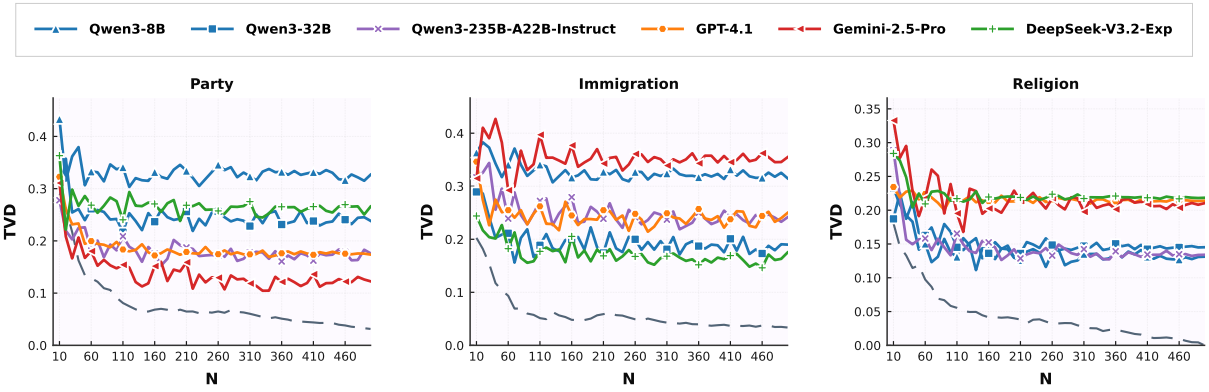


Figure 6: Population simulation reliability versus population coverage. TVD between the aggregated simulated distribution and the target population distribution as a function of the number of simulated individuals N . Curves correspond to different models; the dashed curve shows GT sampling. Lower TVD indicates better population-level agreement. The value of c is 100%.

5 Related Work

5.1 LLM-Based Simulation Evaluation

LLM-based human behavior simulation is increasingly used to generate survey responses, motivating benchmarks that target simulation reliability itself. Santurkar et al. (2023) and Durmus et al. (2024) use real polls to test whether models match human viewpoint distributions across groups and countries. Hu et al. (2025) further treats simulation as a standalone capacity, covering diverse tasks with a unified, comparable protocol. Meister et al. (2024) evaluates 43 LLMs against the American Community Survey and reports label bias and near-random responding, suggesting that the alignment scores may reflect closeness to uniform. However, existing evaluations still emphasize end-to-end scores, offering limited insight into where biases arise and how they propagate.

5.2 Individual-level Simulation

In individual-level simulation evaluation, prior work has explored methods for improving per-respondent simulation accuracy. Bisbee et al. (2024) and Tzachristas et al. (2025) show that zero-shot prompting is easy to deploy but can be highly sensitive to prompt design and sampling randomness, leading to unstable per-respondent accuracy. Wang et al. (2025a) further finds that prompting-based simulation often underestimates the variance of human opinions, producing overly uniform outputs that obscure individual heterogeneity. Beyond prompting, Wang et al. (2024) studies conjoint analysis and proposes a statistical data-augmentation framework that integrates LLM-generated data with a small amount of human data to debias synthetic responses. In parallel, Suh et al. (2025) and Cao et al. (2025) directly fine-tune LLMs on survey data to align token-level probabilities with the empirical distribution. However, these efforts largely conflate reliability with end-to-end improvement and offer limited insight into why simulations fail; SRP instead decomposes simulation into layers to localize error sources and guide targeted fixes.

5.3 Population-level Distribution Alignment

Beyond individual-level accuracy, recent work on pluralistic and distributional alignment treats group-level preference distributions as explicit training targets to better reflect heterogeneous values across users and groups. Chakraborty et al. (2024) targets group-level preference distributions for pluralistic alignment. Melnyk et al. (2024) enforces distributional constraints, Poddar et al. (2024) models user/subgroup modes, and Yao et al. (2025) adds group objectives to prevent minority collapse. SRP instead diagnoses *survey-response* distribution shifts induced and propagated by configuration choices, not distribution matching as training.

6 Conclusion

We introduce the Simulation Reliability Prism (SRP), which models LLM-based human behavior simulation as a three-layer generative process and evaluates reliability at both individual and population levels across model capacity, profile completeness, and population coverage. We run controlled experiments on three survey tasks with eleven LLMs to map how reliability varies across layers and configurations, and to derive directions for more reliable simulation.

Limitations

- Our study focuses on human behavior simulation in survey-style tasks; our conclusions may not directly generalize to more complex settings such as interactive dialogue or multi-agent dynamics.
- We operationalize profile completeness by incrementally incorporating fields from a fixed set of attributes. More principled approaches, such as modeling attribute informativeness or constructing higher-signal user profiles are left for future work.
- We study population coverage in its simplest form by sampling and aggregating independently simulated individuals. More realistic approaches, such as stratified or targeted sampling, modeling heterogeneous participation or incorporating networked interactions are left for future work.

References

Gati Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. *Preprint*, arXiv:2208.10264.

Georg Ahnert, Max Pellert, David Garcia, and Markus Strohmaier. 2025. Extracting affect aggregates from longitudinal social media data with temporal adapters for large language models. *Proceedings of the International AAAI Conference on Web and Social Media*, 19:15–36.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

The World Values Survey Association. 2022. Wvs database.

Marcel Binz, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, Peter Dayan, Can Demircan, Maria K Eckstein, Noémi Éltető, and 1 others. 2025a. A foundation model to predict and capture human cognition. *Nature*, pages 1–8.

Marcel Binz, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, Peter Dayan, Can Demircan, Maria K. Eckstein, Noémi Éltető, Thomas L. Griffiths, Susanne Haridi, Akshay K. Jagadish, Li Ji-An, Alexander Kipnis, Sreejan Kumar, Tobias Ludwig, Marvin Mathony, Marcelo Mattar, and 21 others. 2025b. Centaur: a foundation model of human cognition. *Preprint*, arXiv:2410.20268.

James Bisbee, Joshua D Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M Larson. 2024. Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, 32(4):401–416.

Yong Cao, Haijiang Liu, Arnav Arora, Isabelle Augenstein, Paul Röttger, and Daniel Hershcovich. 2025. Specializing large language models to simulate survey response distributions for global populations. *Preprint*, arXiv:2502.07068.

Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Furong Huang, Dinesh Manocha, Amrit Singh Bedi, and Mengdi Wang. 2024. Maxmin-rlhf: Alignment with diverse human preferences. *Preprint*, arXiv:2402.08925.

Suhyun Cho, Jaeyun Kim, and Jang Hyun Kim. 2024. Llm-based doppelgänger models: leveraging synthetic data for human-like responses in survey simulations. *IEEE Access*.

Eric Chu, Jacob Andreas, Stephen Ansolabehere, and Deb Roy. 2023. Language models trained on media diets can predict public opinion. *Preprint*, arXiv:2303.16779.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

DeepSeek. 2025. Deepseek-v3.2-exp release | deepseek api docs.

Ricardo Dominguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dünner. 2024. Questioning the survey responses of large language models. *Preprint*, arXiv:2306.07951.

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

Sikt ESS ERIC. 2025. Ess11 - integrated file, edition 4.0 | ess - sikt.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Luke Hewitt, Ashwini Ashokkumar, Isaias Ghezae, and Robb Willer. 2024. Predicting results of social science experiments using large language models. *Preprint*.

Tiancheng Hu, Joachim Baumann, Lorenzo Lupo, Nigel Collier, Dirk Hovy, and Paul Röttger. 2025. Simbench: Benchmarking the ability of large language models to simulate human behaviors. *Preprint*, arXiv:2510.17516.

Tiancheng Hu, Yara Kyrychenko, Steve Rathje, Nigel Collier, Sander van der Linden, and Jon Roozenbeek. 2024. Generative language models exhibit social identity biases. *Preprint*, arXiv:2310.15819.

Akaash Kolluri, Shengguang Wu, Joon Sung Park, and Michael S. Bernstein. 2025. Finetuning llms for human behavior prediction in social science experiments. *Preprint*, arXiv:2509.05830.

Benjamin S. Manning, Kehang Zhu, and John J. Horton. 2024. Automated social science: Language models as scientist and subjects. *Preprint*, arXiv:2404.11794.

Nicole Meister, Carlos Guestrin, and Tatsunori Hashimoto. 2024. Benchmarking distributional alignment of large language models. *Preprint*, arXiv:2411.05403.

Igor Melnyk, Youssef Mroueh, Brian Belgodere, Mattia Rigotti, Apoorva Nitsure, Mikhail Yurochkin, Kristjan Greenewald, Jiri Navratil, and Jerret Ross. 2024. Distributional preference alignment of llms via optimal transport. *Preprint*, arXiv:2406.05882.

Suhong Moon, Marwa Abdulhai, Minwoo Kang, Joseph Suh, Widyadewi Soedarmadji, Eran Kohen Behar, and David M. Chan. 2024. Virtual personas for language models via an anthology of backstories. *Preprint*, arXiv:2407.06576.

OpenAI. 2025. Introducing gpt-4.1 in the api | openai.

Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. 2024. Personalizing reinforcement learning from human feedback with variational preference learning. *Preprint*, arXiv:2408.10075.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? *Preprint*, arXiv:2303.17548.

Joseph Suh, Erfan Jahanparast, Suhong Moon, Minwoo Kang, and Serina Chang. 2025. Language model fine-tuning on scaled survey data for predicting distributions of public opinions. *Preprint*, arXiv:2502.16761.

Ioannis Tzachristas, Santhanakrishnan Narayanan, and Constantinos Antoniou. 2025. Guided persona-based ai surveys: Can we replicate personal mobility preferences at scale using llms? *arXiv preprint arXiv:2501.13955*.

Angelina Wang, Jamie Morgenstern, and John P Dickerson. 2025a. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, pages 1–12.

Jia Wang, Ziyu Zhao, Tingjuntao Ni, and Zhongyu Wei. 2025b. Sociobench: Modeling human behavior in sociological surveys with large language models. *Preprint*, arXiv:2510.11131.

Mengxin Wang, Dennis J Zhang, and Heng Zhang. 2024. Large language models for market research: A data-augmentation approach. *arXiv preprint arXiv:2412.19363*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Binwei Yao, Zefan Cai, Yun-Shiuan Chuang, Shanglin Yang, Ming Jiang, Diyi Yang, and Junjie Hu. 2025. No preference left behind: Group distributional preference optimization. *Preprint*, arXiv:2412.20299.

A Theoretical Insight

We provide theoretical insights into when and how the reliability of LLM-based human behavior simulation can converge to its ideal limit. Rather than viewing reliability as a single score, we show that it emerges from a multi-layer structure.

Multi-layer structure Let s denote a task scenario, u an individual, x an individual profile, and y a behavioral outcome. Human behavior is assumed to arise from a structured process. The target behavior distribution under scenario s is

$$p^*(y \mid s) = \sum_u p^*(u) \sum_x p^*(x \mid u)\, p^*(y \mid s, x, u). \tag{8}$$

An LLM-based simulator induces an approximate distribution of the same form,

$$p(y \mid s) = \sum_u p(u) \sum_x p(x \mid u)\, p(y \mid s, x, u), \tag{9}$$

where deviations arise from imperfect modeling of individual sampling, profile construction, and scenario-conditioned behavior generation.

Reliability gap. We define the population-level discrepancy under scenario s as the ℓ_1 distance

$$D(s) \triangleq \left\| p^*(\cdot \mid s) - p(\cdot \mid s) \right\|_1. \tag{10}$$

Layer-wise error propagation. Consider a single (u, x) term and define

$$\begin{aligned} A^* &= p^*(y \mid s, x, u), \\ B^* &= p^*(x \mid u), \\ C^* &= p^*(u). \end{aligned} \tag{11}$$

and their simulator counterparts

$$A = p(y \mid s, x, u), \quad B = p(x \mid u), \quad C = p(u). \tag{12}$$

The discrepancy contributed by this term is $\left|A^*B^*C^* - ABC\right|$. By adding and subtracting intermediate terms and applying the triangle inequality, we have

$$\begin{aligned} |A^*B^*C^* - ABC| \le\ & |B^*C^*|\,|A^* - A| \\ & + |AC^*|\,|B^* - B| \\ & + |AB|\,|C^* - C|. \end{aligned} \tag{13}$$

Bounding the overall error. For any fixed y, the difference between the target and simulated probabilities can be written as

$$p^*(y \mid s) - p(y \mid s) = \sum_u \sum_x \big(A^*B^*C^* - ABC\big). \tag{14}$$

Applying the triangle inequality and summing over y yields

$$\begin{aligned} D(s) &= \sum_y \left|p^*(y \mid s) - p(y \mid s)\right| \\ &\le \sum_{y,u,x} \left|A^*B^*C^* - ABC\right|. \end{aligned} \tag{15}$$

Since all terms are probabilities in $[0, 1]$, we can further upper bound each term by

$$|A^*B^*C^* - ABC| \le |A^* - A| + |B^* - B| + |C^* - C|. \tag{16}$$

which gives the overall bound

$$D(s) \le \sum_y \sum_u \sum_x \Big(|A^* - A| + |B^* - B| + |C^* - C|\Big). \tag{17}$$

Convergence conditions. The bound implies that $D(s) \to 0$ is guaranteed when the following component distributions converge:

$$\begin{aligned} p(u) &\to p^*(u), \\ p(x \mid u) &\to p^*(x \mid u), \\ p(y \mid s, x, u) &\to p^*(y \mid s, x, u). \end{aligned} \tag{18}$$

This result shows that reliable simulation cannot be attained by optimizing any single component in isolation. Instead, reliability emerges from coordinated convergence across individual sampling, profile construction, and scenario-conditioned behavior generation.

B Data Construction & Statistics

To assess LLM-based human behavior simulation across multiple behavioral domains, we run experiments on three survey tasks: *religious stance*, *partisan preference*, and *immigration attitude*. Our task definitions are drawn from commonly used social-science survey resources, including the European Social Survey (ESS; *Party*) (ESS ERIC, 2025), the World Values Survey (WVS; *Immigration*) (Association, 2022), and SocioBench (Socio; *Religion*) (Wang et al., 2025b).

For each dataset, we follow a three-step process. First, we select the survey questions that will define the simulation tasks. Second, we preprocess the raw survey data by merging basic demographic attributes with historical survey responses to construct individual profiles. Finally, we filter the population to retain only respondents with complete and valid profile fields. After filtering, Table 3 reports statistics of the remaining valid data and the associated survey question for each task. Table 4 lists the response options for each survey question.

C Experimental Details

C.1 Prompt Templates For Prediction

Figure 7 shows the prompt template used to elicit survey-style responses in the International Social Survey Programme setting. Our design follows prior prompt-based survey simulation setups in Wang et al. (2025b). Each instance provides (i) an instruction to role-play as a real individual, (ii) the respondent's personal attributes and any previous answers, (iii) the target question and its candidate options, and (iv) an explicit JSON output schema. We require the model to justify its choice in 6–10 sentences based only on the provided attributes, and to return the selected option as a number in the `option` field. This structured format standardizes generation across models and facilitates automatic parsing and evaluation.

C.2 Generation Settings

We employ greedy decoding across all models. For Qwen3-8B, Qwen3-14B, and Qwen3-32B, we ac-

Item	Party	Immigration	Religious
`profile_attributes_total_num`	100	200	80
`population_samples_total_num`	1000	10000	500
`question`	Which political party do you feel closest to? (Germany)	How would you evaluate the impact of immigrants on the development of your country?	Do you think churches and religious organizations have too much or too little power?

Table 3: Task settings and question texts.

Party (options)	Immigration (options)	Religious (options)
1. CDU/CSU	1. Very bad	1. Far too much power
2. SPD	2. Quite bad	2. Too much power
3. The Left (Die Linke)	3. Neither good nor bad	3. About the right amount
4. Alliance 90/The Greens	4. Quite good	4. Too little power
5. FDP	5. Very good	5. Far too little power
6. AFD		
7. Free Voters		
8. dieBasis		
9. Die PARTEI		

Table 4: Answer options for each task.

tivate the `think` mode via `enable_think=True`. In profile-completeness experiments, we evaluate each completeness c across **5 random seeds**, with each seed generating a distinct subset of profile attributes. Similarly, population-coverage experiments use **5 random seeds** for stochastic population sampling. All reported results represent averages over these replications.

D More Experiments

To further test SRP's core claim, we ask whether population-level distribution errors mainly reflect *estimation noise* from limited population coverage or *systematic bias* from insufficient profile information. In the Party scenario, we vary population coverage N while fixing profile completeness c, and measure distributional deviation by TVD. Results are reported in Figure 8.

Across models, TVD decreases rapidly as N increases and then quickly plateaus, indicating that larger N primarily improves stability by reducing variance. In contrast, increasing c produces a more consistent downward shift of the curves, suggesting that the remaining gap is dominated by systematic bias that cannot be removed by sampling more agents alone. Despite differences in absolute performance and smoothness across models, the trend is robust and aligns with SRP: N mainly controls estimation variance, whereas c more directly determines distributional alignment. These results motivate future work on constructing higher-information profiles and explicitly optimizing distribution-level objectives to reduce residual bias.

E Potential Risks

This work examines LLM-based human behavior simulation in survey-style tasks and introduces SRP to diagnose reliability at both the individual and population levels. We use established survey datasets, collect no new personal data, and report results only in aggregate, which helps reduce privacy risks. Still, simulated responses can inherit biases from the underlying data and models, so they should not be interpreted as real public opinion. SRP is meant to support evaluation and auditing, rather than high-stakes use.

F Use of AI Assistant

In this paper, we only use ChatGPT[3] for grammar proofreading and spell checking.

[3] https://chatgpt.com/

Prompt Template

Instruction:
You are participating in the International Social Survey Programme. Assume the role of a real individual with the following personal information. Fully immerse yourself in this persona and answer the question truthfully, based solely on the provided personal information and your previous answers to prior questions.

Personal Information and Previous Answers:
`{attributes}`

Question:
`{question}`

Options:
`{options}`

Please strictly follow the following json format output:

```
{
  "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. Please choose the option that best suits you from the Options given, and respond with the number only.

Figure 7: Prompt template used to conduct simulation.

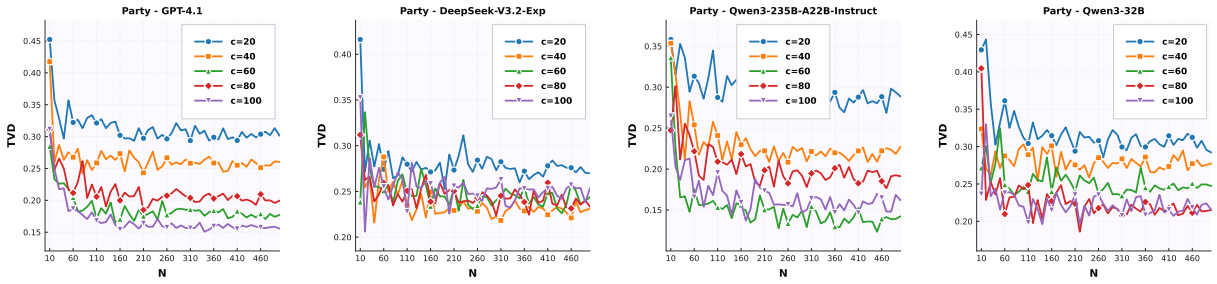


Figure 8: TVD Scaling with Population Coverage N under Varying Profile Completeness c.