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 unreliable simulations can mislead social science conclusions. However, existing evaluations focus on end-to-end scores, leaving it unclear how different aspects of the simulation process interact to determine reliability. We propose ReliMap, which decomposes LLM-based human behavior simulation into three structured layers and evaluates reliability at both the individual level (R1) and population level (R2) across three configuration dimensions: model capacity, profile completeness, and population coverage. Through experiments across four simulation tasks and eleven LLMs, we find that all models exhibit substantial distributional bias without profile conditioning. Profile conditioning reduces this bias with diminishing returns. Larger models benefit more, and attribute informativeness matters more than quantity. Critically, R1 gains do not reliably transfer to R2—individual and population-level reliability can move in opposite directions. At the population layer, increasing coverage reduces variance but not systematic bias, with R2 stabilizing at around 50–100 individuals. These findings highlight that reliable simulation cannot be achieved by optimizing any single layer in isolation, but requires coordinated improvement across all three.[1]

1 Introduction

Human experiments and surveys remain the gold standard for studying behavioral tendencies and social attitudes, yet their scale is fundamentally constrained by cost and participant availability (Aher et al., 2023; Argyle et al., 2023; Bisbee et al., 2024). Large language models (LLMs) have emerged as a promising alternative, enabling rapid, low-cost simulation of human responses at scale.

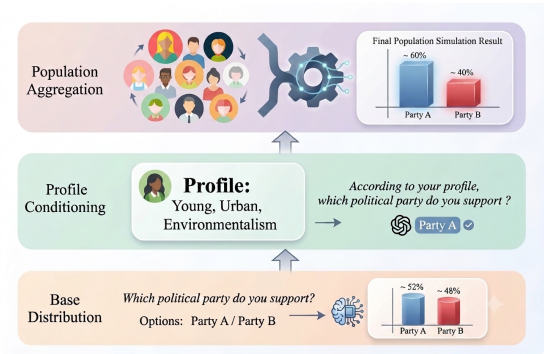


Figure 1: LLM-based Human Behavior Simulation As a Hierarchical Structure.

This has sparked growing interest in using LLMs as computational proxies for human participants (Dominguez-Olmedo et al., 2024; Hu et al., 2024; Manning et al., 2024; Binz et al., 2025; Hu et al., 2025).

Recent years have seen growing efforts in this direction, spanning individual-level response prediction (Aher et al., 2023; Hu et al., 2025; Wang et al., 2025c) and population-level distribution matching (Huang et al., 2025; Wang et al., 2025b). Yet despite this breadth, the field remains fragmented. On the methodology side, some work focuses on prompting strategies and persona engineering to improve per-respondent simulation accuracy (Bisbee et al., 2024; Cho et al., 2024; Moon et al., 2024), while others fine-tune or align models to better reproduce empirical responses (Chu et al., 2023; Suh et al., 2025; Kolluri et al., 2025). On the evaluation side, some work assesses whether simulated individual responses match ground-truth labels (Hewitt et al., 2024), while others examine whether simulated population distributions align with human survey distributions (Huang et al., 2025; Wang et al., 2025b). These lines of work each operate within a narrow scope, addressing one aspect of simulation in isolation. Consequently, it remains unclear how different configuration choices interact to drive reliability. Researchers lack the system-

[1] Our data and code are released at https://anonymous.4open.science/r/simulation_reliability-4D9D/README.md.

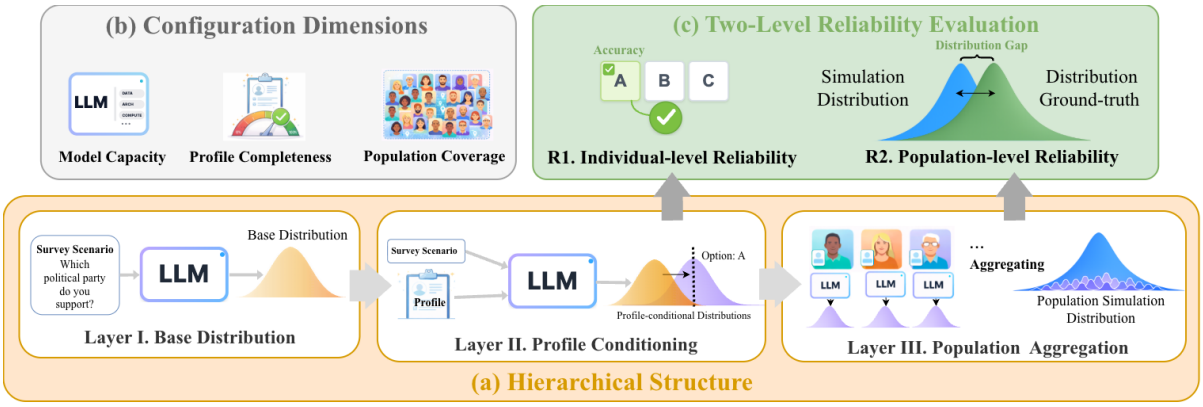


Figure 2: ReliMap Overview: Three-Layer Structure with Configurable Dimensions and Two-Level Reliability Evaluation.

atic understanding needed to diagnose and improve simulation quality.

We argue that this gap stems from treating simulation as a single-step process, when in fact it is hierarchical. Our key insight is that this hierarchical process unfolds in three layers, each building upon the previous one. As illustrated in Figure 1, these three layers are: (1) Base distribution: The model induces an intrinsic behavioral distribution shaped by pretraining and alignment, capturing its default tendency before any individual information is introduced. (2) Profile conditioning: This base distribution is then conditioned on individual profile information—such as demographics, values, or prior experiences—shifting behavioral tendencies toward those of specific personas. (3) Population aggregation: These individualized distributions are finally aggregated across a sampled population, where coverage and sampling strategy determine how faithfully the simulation reflects the target population. Building on this structure, we introduce **ReliMap** to make reliability analysis more structured and interpretable. Specifically, we systematically examine three key configuration dimensions—model capacity, profile completeness, and population coverage—and evaluate their joint effects at both the individual level (R1) and the population level (R2).

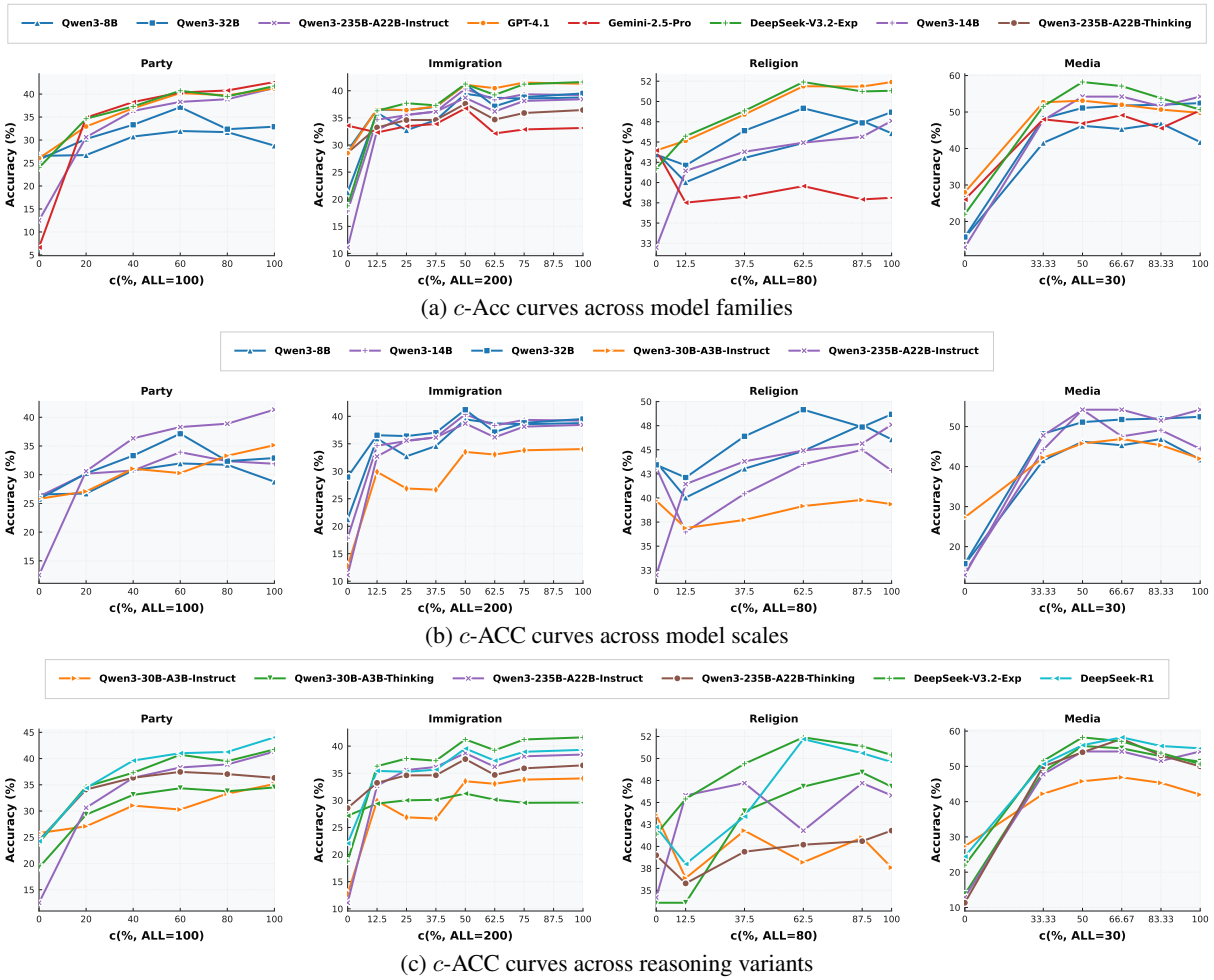
Guided by ReliMap, we conduct experiments on four simulation tasks with eleven LLMs. All models exhibit substantial population-level distributional bias without profile conditioning, with TVD varying widely across models and tasks. Our analysis reveals six findings: (1) Richer profiles improve R1 with diminishing returns. (2) Profile scaling behavior is strongly model-dependent: larger models benefit more from additional profile information. (3) Predictive information is highly concentrated in a small subset of attributes. (4) Attribute informativeness matters more than attribute count. (5) R1 gains do not reliably transfer to R2: individual accuracy improvements can coincide with worsening population-level alignment, revealing a fundamental decoupling between the two levels. (6) Population coverage reduces variance but not systematic bias, with R2 stabilizing at around 50–100 individuals. Together, these findings deepen our understanding of what drives reliability in LLM-based human behavior simulation.

In summary, our contributions are threefold: (1) We introduce ReliMap, which transforms reliability from a single score into a decomposable analysis across three layers, three configuration dimensions, and two evaluation levels, providing a principled framework for understanding and diagnosing LLM-based human behavior simulation. (2) We conduct systematic experiments across four simulation tasks and eleven LLMs, providing a comprehensive empirical examination of how configuration choices shape reliability at each simulation layer. (3) We reveal six findings that fundamentally advance our understanding of simulation reliability, exposing the limitations of current approaches and offering new insights into what it takes to achieve reliable LLM-based human behavior simulation.

2 ReliMap

2.1 Hierarchical Structure

As illustrated in Figure 2(a), we characterize LLM-based human behavior simulation as a three-layer generative process, where each layer corresponds to a natural stage of how simulation is produced.



(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. At each level of c, attributes are sampled randomly.

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, 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

We focus on three configuration dimensions that collectively capture the most critical controllable factors in the simulation setup, rather than

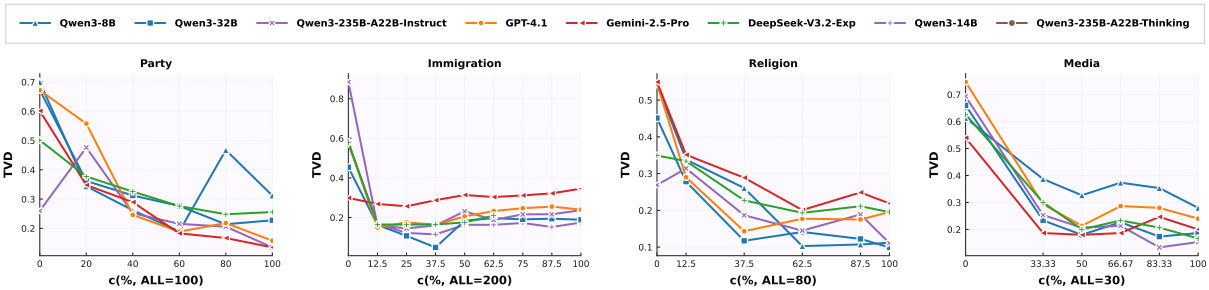


Figure 4: Effect of profile completeness on population simulation reliability.

Model	Party	Immig.	Relig.	Media
Proprietary				
GPT-4.1	0.670	0.368	0.542	0.747
Gemini-2.5-Pro	0.589	0.297	0.549	0.527
Open-source: DeepSeek				
DeepSeek-V3.2-Exp	0.483	**0.128**	0.342	0.613
DeepSeek-R1	0.664	0.579	0.345	**0.480**
Open-source: Qwen3				
Qwen3-8B	0.689	0.304	0.540	0.620
Qwen3-14B	0.722	0.326	0.500	0.578
Qwen3-32B	0.664	0.408	0.474	0.653
Qwen3-30B-A3B-Instruct	0.378	0.396	0.403	0.513
Qwen3-30B-A3B-Thinking	0.527	0.181	0.443	0.647
Qwen3-235B-A22B-Instruct	**0.271**	0.482	0.263	0.684
Qwen3-235B-A22B-Thinking	0.475	0.578	**0.173**	**0.831**

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

implementation-level choices such as prompt templates or decoding strategies. 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 needs 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

As shown in Figure 2(c), we evaluate simulation reliability at two complementary levels: individual-level reliability (R1), which measures per-respondent prediction accuracy, and population-level reliability (R2), which measures how faithfully the aggregated simulation reproduces the true population distribution. These two levels capture distinct aspects of simulation quality and are not interchangeable. Intuitively, stronger R1 should translate into stronger R2, yet this coupling is not guaranteed. We therefore treat their joint evaluation as a necessary condition for a complete assessment of simulation reliability.

Theoretical Insight The three-layer structure, configuration dimensions, and two-level evaluation together form ReliMap. A key theoretical insight underlies this framework: population-level mismatch arises from multiple independent sources across simulation layers, and reducing one source does not automatically compensate for the others. Full derivations are provided in Appendix B.

3 Experiments

3.1 Setup

Data. We conduct experiments on four datasets. The first three are established survey datasets: partisan preference (Party) from the European Social Survey (ESS ERIC, 2025), immigration attitude (Immigration) from the World Values Survey (Association, 2022), and religious stance (Religion) from SocioBench (Wang et al., 2025c). The fourth, Media, differs from the others in two ways. It is a Chinese-language dataset, and it captures a real-world dynamic event—public attitude toward the

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

Luckin Coffee stock collapse—rather than a structured survey. We collect it from a public social media corpus.[3] More details are in Appendix C.

Metric. We evaluate simulation reliability at two levels. For R1, we use accuracy (ACC), which directly measures whether the model's predicted response matches the ground-truth individual behavior. For R2, we use total variation distance (TVD), which quantifies the gap between the simulated and ground-truth population distributions.

Model. We evaluate eleven LLMs covering a diverse range of model families, scales, and reasoning paradigms. Proprietary models include GPT-4.1 (OpenAI, 2025) and Gemini-2.5-Pro (Comanici et al., 2025). Open-source models include DeepSeek-V3.2-Exp (DeepSeek, 2025) and DeepSeek-R1 (Guo et al., 2025), as well as seven models from the Qwen3 family (Yang et al., 2025) spanning different parameter scales and instruct/-thinking variants. This diversity ensures that our findings are not specific to any single model family or architecture.

3.2 Layer I — Base distribution

In Layer I, we examine the model's base behavioral tendency without profile conditioning. We provide no profile information and only input the task context, making one prediction for each instance in the full dataset. 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 TVD of Qwen3-14B and Qwen3-235B-A22B-Instruct differs by approximately 0.45. Moreover, performance shows no consistent correlation with model scale or general capacity, suggesting that population-level distribution alignment constitutes a distinct capacity dimension. Crucially, all models exhibit non-trivial TVD across all tasks, indicating that profile conditioning is necessary to avoid systematic distributional bias.

3.3 Layer II — Profile Conditioning

We investigate the impact of *Model Capacity* and *Profile Completeness*[4] on R1, and obtained the following findings.

Finding 1: More Profile Context Helps, but Saturates Gradually We progressively increase the number of profile attributes and observe the corresponding changes in R1. The results are shown in Figure 3. We find that across four tasks, ACC increases with the number of attributes, but with diminishing marginal gains. This aligns with our intuition that richer user profiles provide models with more personal characteristics, thereby facilitating behavioral simulation. The diminishing returns can be attributed to two factors. On the one hand, newly added attributes may be uninformative or redundant relative to existing ones. On the other hand, processing denser profiles requires stronger contextual understanding and reasoning from the model.

Finding 2: Profile Scaling Behavior is Strongly Model-Dependent To test whether the diminishing returns reflect a capacity bottleneck, we compare how different models scale with profile completeness.

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

[3] Raw data sourced from https://giser2000.github.io/geodata.github.io/#download.

[4] Unless otherwise specified, attributes at each level of c are randomly sampled by default.

Table 2: Attribute informativeness statistics. Gini measures ASPG concentration across attributes (0 = uniform, 1 = maximally concentrated). Top 20% info and Bot 50% info denote the cumulative predictive information contributed by the top 20% and bottom 50% of attributes, respectively. Info@c=50% reports cumulative information captured at half profile completeness under random (rand.) and informativeness-weighted (wtd.) sampling.

	Party	**Immigration**	**Religion**
# attributes	100	200	76
ASPG median	0.017	0.010	0.014
Gini	0.511	0.503	0.443
Top 20% info	53.5%	53.3%	49.7%
Bot 50% info	15.0%	16.9%	21.7%
Info@$c = 50\%$ (rand.)	45%	44%	47%
Info@$c = 50\%$ (wted.)	86%	83%	79%

Model		**Party**	**Immigration**	**Religion**
GPT4.1	r_Q	0.9177	0.8368	**0.9604**
	r_I	0.9928	0.9662	0.9607
	Δ	**+0.0751**	**+0.1294**	**+0.0003**
Gemini-2.5-Pro	r_Q	0.7835	-0.0386	-0.5336
	r_I	0.9624	0.0539	-0.7305
	Δ	**+0.1789**	**+0.0925**	-0.1969
DeepSeek-V3.2-Exp	r_Q	0.8708	0.7233	0.8830
	r_I	0.9885	0.9251	0.9786
	Δ	**+0.1177**	**+0.2018**	**+0.0956**
Qwen3-235B-A22B	r_Q	0.8562	0.6770	0.8485
	r_I	0.9898	0.9082	0.9654
	Δ	**+0.1336**	**+0.2312**	**+0.1169**
Qwen3-32B	r_Q	0.6508	0.7117	0.8588
	r_I	0.8572	0.8911	0.8446
	Δ	**+0.2064**	**+0.1794**	-0.0142
Qwen3-8B	r_Q	0.6016	0.7355	0.8025
	r_I	0.7590	0.9132	0.6217
	Δ	**+0.1574**	**+0.1777**	-0.1808
Mean	r_Q	0.7801	0.6374	0.6366
	r_I	0.9250	0.8126	0.6067
	Δ	**+0.1449**	**+0.1752**	-0.0299

Table 3: Pearson correlations between accuracy and two profile completeness metrics. r_Q: quantity-based completeness; r_I: information-based completeness. $\Delta = r_I - r_Q$.

requires understanding social contexts, individual traits, and behavioral patterns that are potentially distinct from formal reasoning skills.

Finding 3: Behavioral Information Is Highly Concentrated Among Few Attributes To examine why performance saturates as profile completeness increases, we analyze the predictive information carried by individual attributes. We define *Attribute-Specific Predictive Gain (ASPG)* as

$$\text{ASPG}(a) = \frac{I(a; Y)}{H(Y)}, \tag{4}$$

where $I(a; Y)$ is the mutual information between attribute a and target label Y, and $H(Y)$ is the label entropy. ASPG measures how much uncertainty about the target behavior can be reduced by observing a single attribute.[5]

As shown in Table 2, ASPG distributions are strongly right-skewed across all tasks. The top 20% of attributes contribute roughly half of the total predictive information, while the bottom 50% contribute less than 22%. Under the same completeness budget ($c = 50\%$), informativeness-weighted sampling captures 79–86% of total information, compared to only 44–47% under random sampling. These results show that predictive information is unevenly distributed across attributes, highlighting the importance of attribute selection.

Finding 4: Attribute Informativeness Matters More Than Attribute Count Table 3 compares two ways of measuring the correlation between profile completeness and ACC: r_Q measures the Pearson correlation between the proportion of provided attributes and ACC, and r_I measures the Pearson correlation between the cumulative ASPG of provided attributes and ACC. Across most models and tasks, r_I consistently exceeds r_Q. On average, information-based completeness improves the correlation with accuracy by **+0.1449** on *Party* and **+0.1752** on *Immigration*, indicating that the predictive information carried by attributes matters more than their count.

The trend is particularly strong for larger models such as GPT-4.1 and DeepSeek-V3.2-Exp, which achieve near-perfect correlations between r_I and accuracy on several tasks. In contrast, the improvement is weaker or negative on *Religion*, suggesting that the usefulness of informative attributes depends on task characteristics. Although ASPG relies on ground-truth labels and may not be directly applicable in real-world deployment, the results demonstrate that attribute quality is generally more important than attribute quantity.

3.4 Layer III — Population Aggregation.

In Layer III, we evaluate whether improvements in individual-level simulation accuracy from Layer II reliably propagate to population-level distributional alignment. We analyze two questions: (i) how

[5] ASPG analysis excludes the Media task, which uses unstructured historical posts rather than predefined attributes.

Table 4: Comparison of model performance across four datasets with varying profile completeness c. Purple cells indicate ACC↑ and TVD↑; Green cells indicate ACC↓ and TVD↓.

Dataset	GPT-4.1		DeepSeek-V3.2-Exp		Qwen3-235B-A22B	
	ΔACC	ΔTVD	ΔACC	ΔTVD	ΔACC	ΔTVD
Party						
20	+26.50	-55.52	+44.45	-43.66	+144.57	+5.91
40	+12.08	-13.78	+7.68	-16.99	+18.68	-23.62
60	+8.94	-31.88	+9.17	+6.33	+5.39	-39.73
80	-1.19	+14.64	-3.00	-1.33	+1.51	+42.54
100	+3.72	-22.73	+5.67	+5.12	+6.28	-18.26
Immig.						
12.5	+28.51	-56.24	-3.76	+15.78	+194.15	-77.47
25	-0.52	+3.27	+3.56	+2.90	+8.84	+10.58
37.5	+1.87	-13.44	+1.26	-1.08	+1.57	-4.74
50	+10.76	-19.93	+8.48	-40.87	+7.06	-15.16
62.5	-1.53	+24.17	-12.58	+65.56	-6.49	+42.34
75	+2.55	-3.15	+2.30	-3.15	+5.34	-5.56
87.5	-1.74	+13.42	+2.28	+3.07	+1.31	-5.46
100	+1.25	-3.75	-1.43	+3.15	-0.47	+7.48
Relig.						
12.5	+2.73	-56.51	+9.59	-29.64	+29.50	-6.96
37.5	+7.17	-25.37	+6.82	-15.12	+5.69	-34.86
62.5	+7.19	-3.18	+7.29	+1.96	+2.56	+19.91
87.5	-0.08	+9.62	-2.21	-0.76	+1.60	-16.73
100	+1.08	+4.71	+0.16	-2.52	+4.29	-10.54
Med.						
33.3	+88.10	-67.85	+134.34	-58.70	+270.69	-65.91
50	+0.84	+2.77	+12.93	-21.92	+13.49	-10.48
66.7	-2.09	-4.50	-1.91	+3.36	+0.00	-17.02
83.3	-2.56	-7.54	-5.84	-15.22	-4.92	-20.51
100	-1.75	+8.16	-5.79	+2.56	+5.17	+1.60

Table 5: Option-level prediction distributions for two decoupling cases (Qwen3-235B-A22B). **Case 1** (Immigration): c=12.5% →25%, ACC↑, TVD↑. **Case 2** (Media): c=66.7% →83.3%, ACC↓, TVD↓.

Option	GT%	Pred%	Bias	Pred%	Bias
Case 1: Immigration (c=12.5% →25%)					
Very bad	7.49	5.81	-1.68	5.48	-2.01
Quite bad	14.80	30.47	+15.67	23.49	+8.70
Neither	35.71	21.97	-13.74	23.32	-12.39
Quite good	31.17	35.17	+4.00	44.23	+13.06
Very good	10.83	6.57	-4.25	3.47	-7.35
Case 2: Media (c=66.7% →83.3%)					
Loyal Advocate	30.00	16.00	-14.00	18.44	-11.56
Opportunist	26.67	24.89	-1.78	26.89	+0.22
Moral Outrage	12.00	12.22	+0.22	11.33	-0.67
Rational Evaluator	6.00	8.89	+2.89	9.11	+3.11
Bystander Entertainer	25.33	38.00	+12.67	34.22	+8.89

profile completeness gains in Layer II translate to R2(Finding 5), and (ii) how population coverage independently affects R2(Finding 6).

Finding 5: R1 Gains Do Not Reliably Transfer to R2 We track R2 as Layer II profile completeness c increases in Figure 4. R2 improves with c but with diminishing returns: as profile completeness increases, the aggregated distribution moves closer to the empirical one, but the marginal gain shrinks as c grows large and R1 begins to saturate. This overall trend, however, reflects only the average trajectory and does not guarantee that each incremental step in c produces consistent gains at both levels.

This overall improvement trend does not mean R1 gains reliably transfer to R2. As shown in Table 4, ΔACC and ΔTVD can move in opposite directions across tasks, models, and completeness levels. We note that 26% of incremental steps in Table 4 show R1 and R2 moving in opposite directions.

To illustrate how this decoupling manifests in practice, we examine two representative cases in Table 5. In Case 1 (Immigration, $c = 12.5\% \rightarrow 25\%$), ACC improves but TVD worsens: as more profile information is provided, the model increasingly concentrates its predictions on *Quite good* (bias: $+4.00\% \rightarrow +13.06\%$) while further under-predicting *Neither good nor bad* and *Very good*, pushing the predicted distribution away from the true population distribution. In Case 2 (Media, $c = 66.7\% \rightarrow 83.3\%$), the reverse holds: the model's over-prediction of *Bystander Entertainer* and under-prediction of *Loyal Advocate* both partially correct, bringing the predicted distribution closer to the truth, but individual accuracy drops. Together, the cases reveal that TVD is driven by distributional shifts in a small subset of options, which can move independently of overall ACC.

Finding 6: Population Coverage Reduces Variance but Not Bias 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 5. This distinction between variance and bias is further corroborated in Appendix F, where we jointly vary N and c.

Results show that 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. When N is large, the remaining error is driven primarily by systematic distributional mismatch rather than sampling noise. 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

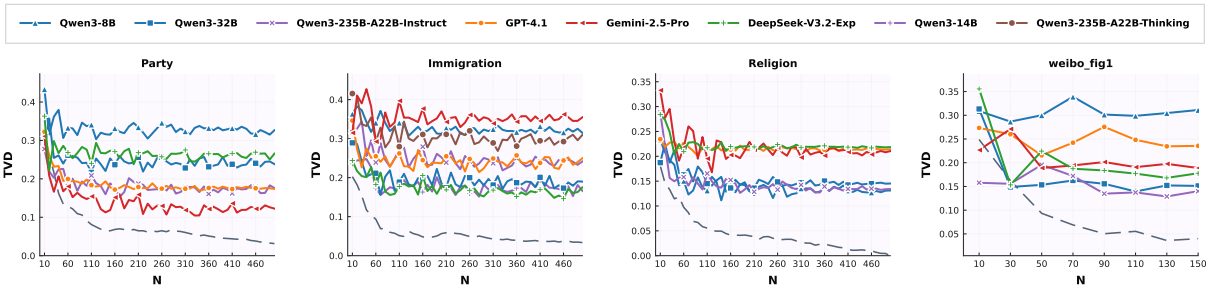


Figure 5: Population simulation reliability versus population coverage. TVD between the aggregated simulated distribution and the target population distribution as a function of the number of simulated individuals N. The dashed curve shows GT sampling which denotes human ground-truth responses aggregated at varying sampling sizes, serving as a reference for systematic sampling error. The c is fixed at 100%.

sampling variance but cannot eliminate the residual distributional mismatch.

4 Related Work

4.1 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.

4.2 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; ReliMap instead decomposes simulation into layers to localize error sources and guide targeted fixes.

4.3 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. ReliMap instead diagnoses *survey-response* distribution shifts induced and propagated by configuration choices, not distribution matching as training.

5 Conclusion

We introduce ReliMap, 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 four simulation 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 evaluate individual-level reliability using accuracy, which directly reflects the discrete nature of survey responses but is a winner-takes-all metric that does not capture the subjectivity inherent in human decision-making. More principled alternatives remain an open direction for future work.

- Our analysis is observational. While we identify how configuration choices affect reliability, the underlying mechanisms—such as why larger models benefit more from richer profiles—remain unexplored. Interpretability of LLM-based simulation is an important direction 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.

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. 2025. A foundation model to predict and capture human cognition. *Nature*, pages 1–8.

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.

Ji Huang, Mengfei Li, and Shuai Shao. 2025. Distribution shift alignment helps llms simulate survey response distributions. *Preprint*, arXiv:2510.21977.

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.

Han Wang, Jacek Pawlak, and Aruna Sivakumar. 2025b. Can large language models simulate human responses? a case study of stated preference experiments in the context of heating-related choices. *arXiv preprint arXiv:2503.10652*.

Jia Wang, Ziyu Zhao, Tingjuntao Ni, and Zhongyu Wei. 2025c. 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 Potential Risks

This work examines LLM-based human behavior simulation in survey-style tasks and introduces ReliMap 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. ReliMap is meant to support evaluation and auditing, rather than high-stakes use.

B 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

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

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

$$p(y \mid s) = \sum_u p(u) \sum_x p(x \mid u)\, p(y \mid s, x, u), \tag{6}$$

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

$$D(s) \triangleq \left\| p^*(\cdot \mid s) - p(\cdot \mid s) \right\|_1. \tag{7}$$

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

$$\begin{aligned} A^* &= p^*(y \mid s, x, u), \\ B^* &= p^*(x \mid u), \\ C^* &= p^*(u). \end{aligned} \tag{8}$$

and their simulator counterparts

$$A = p(y \mid s, x, u), \quad B = p(x \mid u), \quad C = p(u). \tag{9}$$

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

$$\begin{aligned} |A^* B^* C^* - ABC| \leq\ & |B^* C^*|\,|A^* - A| \\ & + |AC^*|\,|B^* - B| \\ & + |AB|\,|C^* - C|. \end{aligned} \tag{10}$$

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

$$p^*(y \mid s) - p(y \mid s) = \sum_u \sum_x \left(A^* B^* C^* - ABC \right). \tag{11}$$

Applying the triangle inequality and summing over y yields

$$\begin{aligned} D(s) &= \sum_y \left| p^*(y \mid s) - p(y \mid s) \right| \\ &\leq \sum_{y,u,x} \left| A^* B^* C^* - ABC \right|. \end{aligned} \tag{12}$$

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

$$|A^* B^* C^* - ABC| \leq |A^* - A| + |B^* - B| + |C^* - C|. \tag{13}$$

which gives the overall bound

$$D(s) \leq \sum_y \sum_u \sum_x \Big(|A^* - A| + |B^* - B| + |C^* - C| \Big). \tag{14}$$

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

$$\begin{aligned} p(u) &\to p^*(u), \\ p(x \mid u) &\to p^*(x \mid u), \\ p(y \mid s, x, u) &\to p^*(y \mid s, x, u). \end{aligned} \tag{15}$$

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.

C Data Construction Statistics

We conduct experiments on four tasks across two settings. The first setting comprises three established survey tasks: *partisan preference*, *immigration attitude*, and *religious stance*. Task definitions are drawn from the European Social Survey (ESS; *Party*) (ESS ERIC, 2025), the World Values Survey (WVS; *Immigration*) (Association, 2022), and SocioBench (*Religion*) (Wang et al., 2025c). For each dataset, we follow a three-step process. First, we select the survey questions that define the simulation tasks. Second, we preprocess the raw survey data by merging 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. Table 6 reports statistics of the remaining valid data and the associated survey question for each task. Table 7 lists the response options for each survey question.

The second setting introduces a real-world dynamic event: public attitude toward the Luckin Coffee stock collapse. It is in Chinese, further testing whether our findings generalize across languages and beyond traditional survey scenarios. Raw data were sourced from a publicly available social media dataset.[6] We filtered it for original posts containing the hashtag "#Luckin Coffee Collapse" during the event period. Ground-truth labels were obtained by applying GPT-4o as an annotator to perform sentiment analysis on users' original posts. Individual profiles were constructed from each user's 30 most recent historical posts prior to the event. Table 6 includes statistics for this task alongside the three surveys.

[6] https://giser2000.github.io/geodata.github.io/#download.

Item	Party	Immigration	Religious	Media
`profile_attr_num`	100	200	80	30
`population_num`	1,000	9,994	500	150
`question`	Which political party do you feel closest to? (Germany)	How would you evaluate the impact of immigrants on your country's development?	Do churches and religious organizations have too much or too little power?	Analyze this user's attitude toward the 'Luckin Coffee stock price crash' incident.

Table 6: Task settings and question texts.

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

Table 7: Answer options for each task.

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}`

JSON Output:

```
{"reason": "", "option": ""}
```

Requirements:

1. Answer based on personal information only; 6–10 sentence justification.
2. Choose best option; respond with number only.

Figure 6: Prompt template for Party, Immigration, Religion.

Instruction: You are a user attitude simulation model. Based on historical Weibo posts provided, simulate this user's likely attitude toward "Luckin Coffee's stock price plummeted" and determine which category best matches.

Five Attitude Categories:

1. **Loyal Advocacy:** High emotional support; views Luckin as "pride of domestic brands"; heartbreak/defensiveness on stock crash without blaming company.
2. **Opportunistic Exploitation:** No loyalty; only cares about prices/discounts/investment; neutral on fraud; motivation: "getting a bargain" or "buying the dip."
3. **Moral Indignation:** Strong moral outrage on corporate integrity; challenges "national brand" narrative; fraud is unforgivable.
4. **Belief Evaluation:** Distinguishes product from corporate behavior; may acknowledge quality but does not forgive fraud; rational, analytical stance.
5. **Entertainment Spectator:** No substantive stance; memes/jokes/viral quips (e.g., "Sauce-flavored Latte becomes Limit-down Latte"); lighthearted, social interaction focused.

Input: `{posts}` → **Event:** Luckin Coffee's stock price plummeted

Output Rules:

1. Output exactly one valid JSON object only, no extra text/spaces/line breaks/Markdown.
2. JSON must contain: `analysis` (string) and `attitude_id` (integer 1–5).
3. Do NOT use "`` `json or code block markers. Output raw JSON directly.

JSON Output:

```
{"analysis": "", "attitude_id": 0}
```

Figure 7: Prompt template for Media Task (translated from Chinese).

D Experimental Details

D.1 Prompt Templates For Prediction

Figure 6 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. (2025c). 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. The prompt design for the Media task follows a similar structure, as shown in Figure 7.

D.2 Generation Settings

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

E Illustrative Example of the Simulation Pipeline

To clarify how the formal model is instantiated in practice, we provide a concrete example from the Party task. Table 8 shows how the three key components—scenario s, individual u, and profile x—are populated in a real simulation instance.

In this framework, u refers to a specific survey respondent (e.g., Respondent_1). The scenario s defines the task context, including the question and candidate options. The profile x is a subset of attributes sampled from the full set of attributes available for u, which corresponds to $p(x \mid u)$ in our formulation. Studying how simulation performance varies as we include more or fewer attributes from x allows us to quantify the effect of profile completeness on reliability.

Note that two users with identical attribute sets x will indeed generate identical prompts. In practice, however, since we employ a rich set of high-dimensional sociodemographic features, cases where two distinct individuals share the same u are extremely rare in our dataset and can be considered negligible.

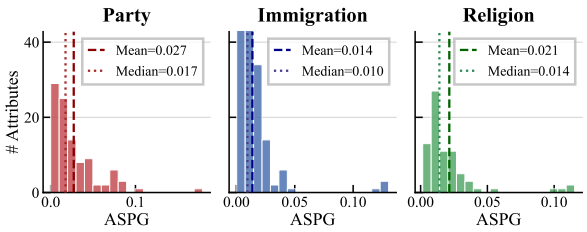


Figure 8: Distribution of ASPG values across attributes for each task. Most attributes carry low predictive information individually, while a small minority contributes disproportionately large predictive gain.

F More Experiments

F.1 ASPG Distribution

Figure 8 shows the distribution of ASPG values across attributes for each task. In all three tasks, the distribution is strongly right-skewed: most attributes carry low predictive information individually, while a small minority contributes disproportionately large predictive gain.

F.2 Effect of Population Coverage and Profile Completeness on Distributional Bias

To further test ReliMap'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 9.

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 ReliMap: 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.

Component	Content
s (scenario)	*Question*: Which political party do you feel closest to? (Germany) *Options*: 1. CDU/CSU, 2. SPD, 3. The Left (Die Linke), 4. Alliance 90/The Greens, 5. FDP, 6. AFD, 7. Free Voters, 8. dieBasis, 9. Die PARTEI
u (individual)	Respondent_1
x (profile)	{"What the country needs most is loyalty towards its leaders.": "Disagree strongly", "How worried are you about climate change?": "Not at all worried", "Have you felt unfairly treated by the police because you are a man/woman?": "No", ...}
Prompt	`### 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:` {content from x} `### Question:` {content from s} `### Options:` {content from s}
Response	{"reason": "Based on my previous answers, I am strongly in favor of gender equality in business leadership [...] SPD is also traditionally supportive of worker rights and moderate reforms, which matches my practical, balanced approach. Therefore, I feel closest to the SPD.", "option": "2"}
Ground Truth	2
ACC	1.0

Table 8: A concrete example of the simulation pipeline for the Party task, illustrating how s, u, and x are instantiated in practice.

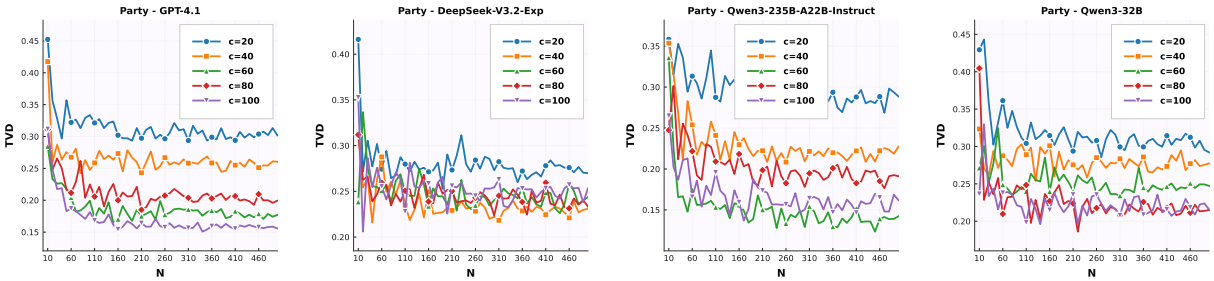


Figure 9: TVD Scaling with Population Coverage N under Varying Profile Completeness c.

G Robustness Analysis

To examine whether our findings are sensitive to implementation-level choices, we conduct robustness tests on prompt design and decoding temperature using the Party task across three models. Each setting is evaluated over 5 independent sampling runs, with results reported as mean individual-level simulation accuracy (ACC).

G.1 Prompt Variations

We test two additional prompt variants alongside the original prompt described in Appendix C.

Original prompt follows the role-playing instruction described in Appendix C, asking the model to fully immerse itself in the respondent's persona.

Prompt_v1 instructs the model to predict the respondent's most probable choice instead of role-playing: *"You are analyzing responses for the International Social Survey Programme. Given the demographic and attitudinal profile below, predict the most likely response this individual would provide to the survey question."*

Prompt_v2 directly instructs the model to adopt the role of the respondent with a minimal instruction: *"You're taking part in an international survey. Answer as a real person with these traits:"*

G.2 Temperature

We test decoding temperature $t = 0.6$ alongside the default greedy decoding ($t = 0$) using the original prompt template.

Across all prompt variants and temperature settings, two patterns remain consistent. First,

Model	Prompt	Temperature	c (%)					
			0	20	40	60	80	100
Qwen3-8B	Original	greedy	26.53	26.72	30.76	31.96	31.72	28.78
	Prompt_v1	greedy	28.40	31.00	32.57	34.16	32.08	33.26
	Prompt_v2	greedy	26.40	28.07	29.81	31.29	30.81	31.95
	Original	$t = 0.6$	26.52	27.22	31.05	32.88	31.87	32.53
Qwen3-32B	Original	greedy	25.76	30.32	33.32	37.14	32.26	32.90
	Prompt_v1	greedy	27.34	33.08	35.02	38.70	36.32	37.78
	Prompt_v2	greedy	26.38	30.90	33.26	35.08	32.91	33.68
	Original	$t = 0.6$	26.00	30.52	33.28	35.54	33.38	32.84
Qwen3-235B-A22B-Instruct	Original	greedy	12.52	30.62	36.34	38.30	38.88	41.32
	Prompt_v1	greedy	21.04	32.72	35.98	39.68	38.76	42.78
	Prompt_v2	greedy	18.78	29.28	35.20	37.44	38.40	39.66
	Original	$t = 0.6$	11.25	29.10	33.10	38.84	37.96	38.18

Table 9: Individual-level simulation accuracy (ACC, %) under different prompt variants and decoding temperatures on the Party task. Original refers to the default prompt used in the main experiments.

individual-level simulation accuracy improves with profile completeness but with diminishing marginal gains, confirming the robustness of Finding 1. Second, larger models achieve higher accuracy and continue to benefit more from richer profiles, confirming the robustness of Finding 2. While absolute accuracy varies across settings, the macro-level patterns identified in our main experiments remain stable, suggesting that our findings are not artifacts of particular implementation choices.

H Statistical Uncertainty Analysis

To assess the statistical reliability of our experimental findings, we report the mean and standard deviation of the individual-level accuracy (R1) across 5 fixed random seeds (2025, 2026, 2027, 2028, 2029), where randomness stems from the attribute subset sampling procedure.

Figure 10 replicates Figure 3(a) across all four tasks (Party, Immigration, Religion, and Media) with ± 1 standard deviation shaded regions. The shaded regions reflect the variability introduced by different random attribute subset samples rather than model stochasticity. As shown, while some tasks (e.g., Religion) exhibit larger variance due to the smaller dataset size, the main trends remain consistent across models and tasks, supporting the reliability of our conclusions.

I Use of AI Assistant

In this paper, we only use ChatGPT[7] for grammar proofreading and spell checking.

[7] https://chatgpt.com/

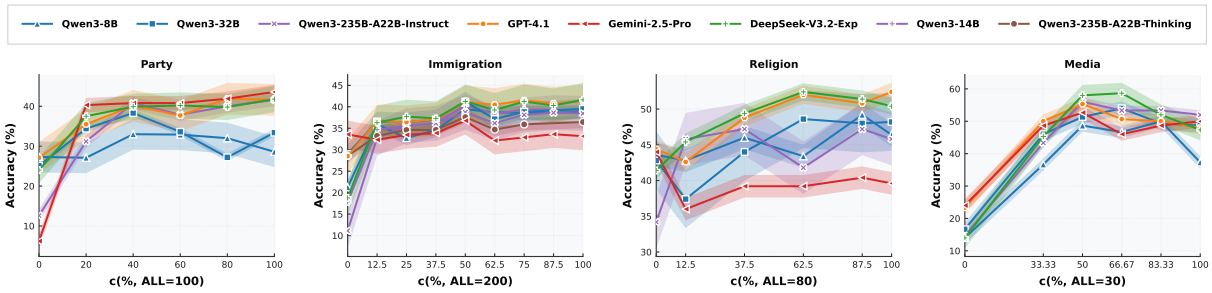


Figure 10: Individual-level accuracy (R1) as a function of profile completeness c across all four tasks, with ± 1 standard deviation shaded regions computed over 5 fixed random seeds (2025–2029). Variance stems from attribute subset sampling. Despite variability in some tasks, the overall trends remain consistent across models.