Context Modeling for LLM-Based Behavioral Simulation

Anonymous ACL submission

Abstract

Large language models (LLMs) offer new opportunities for simulating human decision making at scale, yet prompting-based approaches often fail to recover realistic population-level response distributions, particularly for heterogeneous demographic groups and high-entropy questions. This paper introduces context modeling with agentic AI, a structured simulation framework that explicitly constructs latent, individual-level decision contexts prior to response generation. By separating persona construction from decision making and introducing structured stochasticity through context sampling, the method captures within-group heterogeneity and mitigates mode collapse inherent in single-prompt simulations. Extensive empirical evaluation across population-level and group-conditioned settings shows that context modeling consistently improves simulation fidelity, with especially strong gains on high-entropy questions and demographic subpopulations, while preserving performance on low-entropy tasks. Detailed tail analysis further reveals that the primary improvement arises from recovering minority and low-probability response options that are systematically underrepresented by direct prompting. Overall, the results demonstrate that faithful behavioral simulation requires explicit modeling of contextual heterogeneity and structured uncertainty, establishing context modeling as a robust foundation for LLM-based behavioral simulation.

1 Introduction

The ability to simulate human decision-making has long been a central goal across the social sciences, economics, psychology, and public policy (Sun, 2018; Chen et al., 2024). Traditional simulation methods such as agent-based models, system dynamics, econometric formulations, and survey-driven statistical techniques have yielded valuable insights, but they rely on predefined assumptions about which variables matter (Abar et al., 2017).

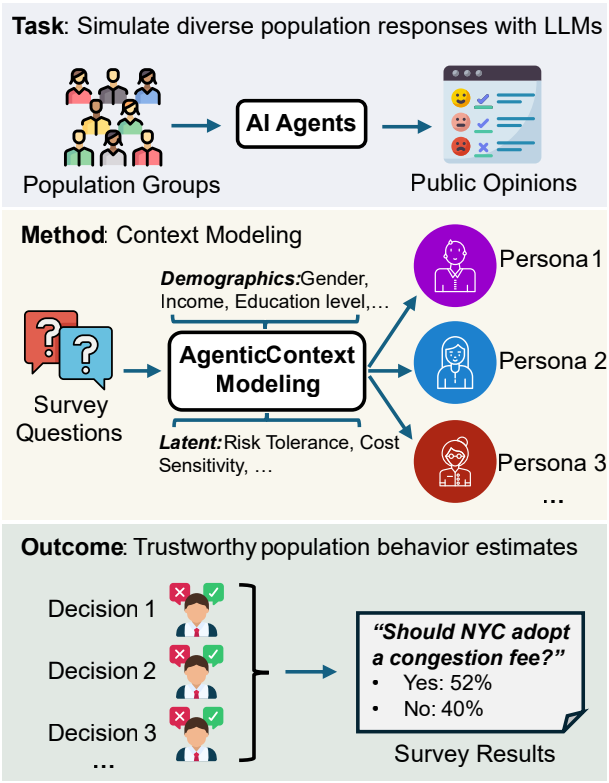

Figure 1: Conceptual overview of context modeling for population-level behavior simulation with LLMs.

These assumptions limit scalability and can fail to capture the richness and complexity of human reasoning. Recent advances in large language models (LLMs) have introduced a new paradigm for behavioral simulation (Gurcan, 2024). Unlike traditional models that require explicit specification of features, LLMs can draw on broad latent knowledge to produce responses that are coherent, context-sensitive, and normatively grounded (Qiu et al., 2025). Prior work on LLM-based behavioral simulation has revealed systematic limitations, including insufficient stochasticity (Lior et al., 2025), over-reliance on single personas, low performance on high-entropy questions (open questions) (Tonmoy et al., 2024), and limited ability to

faithfully represent specific groups (Simmons and Hare, 2023). To address this challenge, this paper introduces context modeling (Figure 1), an agentic framework that explicitly constructs individual-level decision contexts by identifying and sampling contextual factors before generating simulated responses with LLMs.

A key motivation for context modeling lies in its potential to capture *"unknown unknowns"* (Beigi et al., 2024). Human-decisions are shaped not only by observable attributes such as age, income, and education, but also by latent factors including risk preferences, norms, personal experience, and situational constraints (Asparouhov, 2005). Traditional simulation approaches assume that all relevant determinants are known in advance, an assumption rarely satisfied in practice (Stern, 1997). By contrast, context modeling reframes LLM-based behavioral simulation as a conditional text generation problem under latent contextual uncertainty, and studies how structured context conditioning reshapes the induced output distribution.

Building on this design, this paper elaborates context modeling for LLM-based behavioral simulation through the following research questions:

RQ1: Can context modeling improve LLM-based behavioral simulation over direct prompting and numerical baselines?

RQ2: What mechanisms enable context modeling to improve distributional fidelity in LLM-based simulation?

RQ3: Does context modeling accurately represent specific demographic groups and improve simulation performance for those groups?

RQ4: Can context modeling enhance stochasticity for high-entropy questions (open questions) while performing well on low-entropy (concensus) questions?

2 Literature Review

2.1 LLMs for Behavioral Simulation

LLM-based simulation has been widely explored for modeling and mimicking human decision-making (Lee et al., 2025; Light et al., 2025). The underlying idea is that, when appropriately prompted, an LLM can approximate how individuals from a target population respond to survey questions, moral dilemmas, or policy scenarios(Wang et al., 2025). If reliable, such simulations could complement or partially replace traditional surveys, enabling rapid and low-cost exploration of behavioral responses across diverse populations (Anthis et al., 2025). Early studies reported promising results, showing that LLM-generated responses can resemble aggregate human statistics in domains such as opinion surveys and moral reasoning, and that LLM-based agents can produce plausible social behaviors in interactive settings (Park et al., 2024; Binz and Schulz, 2023).

However, subsequent work has exposed significant limitations. LLM responses are often highly sensitive to prompt wording, persona descriptions, and contextual framing, leading to instability across runs (Breznau et al., 2022). More importantly, asking a model to respond as a "typical" member of a group collapses within-group variation into a single deterministic representation, failing to reflect the diversity of real populations (Clinton et al., 2025). Some studies also prove that even state-of-the-art models achieve limited fidelity to human response distributions, particularly on high-entropy questions where opinions are diverse (Hu et al., 2025). These findings suggest that naive prompting alone is insufficient for faithful behavioral simulation.

2.2 Persona Prompting and Demographic Conditioning

Persona prompting is a common technique for conditioning LLMs on demographic or role-based attributes, instructing the model to answer from the perspective of a specified individual or group (Tissaoui, 2025; Amin et al., 2025). This approach can guide model outputs toward subgroup-specific behaviors and has been applied in domains ranging from ethical reasoning to survey simulation (Kim et al., 2024). In some cases, persona conditioning improves alignment with known group-level trends compared to unconditioned prompting (Salminen et al., 2022).

Despite its usefulness, persona prompting has fundamental limitations. It depends on the model's internal representations of demographic categories, which may be biased or overly stereotypical, and a single static persona cannot capture the heterogeneity present within real groups (Tissaoui, 2025; Liu et al., 2023). As a result, persona-based simulations often over-smooth responses and underrepresent minority viewpoints, especially for questions with diverse opinions (Amin et al., 2025). Recent efforts have attempted to address this issue by sampling multiple personas or applying post-hoc calibration, but these methods lack a principled mechanism for modeling latent contextual factors (Salminen et al.,

2020; Lutz et al., 2025). This limitation motivates approaches that explicitly represent and sample individual-level contexts as a foundation for more faithful behavioral simulation.

3 Problem Definition and Notation

Consider a behavior question $q \in \mathcal{Q}$ with answer options $\mathcal{A}_q$ and demographic groups $g \in \mathcal{G}$. Human responses for each question–group pair (q, g) define an empirical distribution $p^{\mathrm{H}}(a \mid q, g)$, estimated from survey data. Given a simulation method m and an LLM ℓ, the simulator induces a corresponding distribution $p^{\mathrm{S}}_{m,\ell}(a \mid q, g)$. The objective of LLM-based behavioral simulation is to design methods for which $p^{\mathrm{S}}_{m,\ell}$ closely approximates p^{H} across diverse questions and groups, ensuring robust population-level and group-level simulation.

4 Agentic Context Modeling

In context modeling, each individual answering question q within group g is assumed to possess an unobserved context vector $c \in \mathcal{C}$ that represents latent characteristics influencing the individual's decision. Let $c = (c_1, \ldots, c_M)$ denote the components of these characteristics (e.g., age, gender, income). The simulation framework approximates the human response distribution by repeatedly sampling such context vectors from a distribution $p(c \mid q, g)$ and conditioning an LLM prompt on each sampled c to generate a predicted answer a. At a conceptual level, the context modeling simulation pipeline consists of a controller and four working agents (Figure 2):

1. **Feature Selection**: Identify a set of context variables that are relevant for predicting responses to question q (Section 4.1).

2. **Distribution Estimation**: For each selected variable, use an LLM agent to estimate a distribution $p(C_j \mid q, g)$ representing plausible values that individuals in group g may take for that variable (Section 4.2).

3. **Persona Sampling**: Sample N independent personas $c^{(1)}, \ldots, c^{(N)}$ from the joint distribution $p(c \mid q, g)$, where each persona encodes a concrete instantiation of the selected context variables (Section 4.3).

4. **Decision Generation**: For each sampled persona, construct a prompt that incorporates both the persona description and the question q, query the LLM for an answer, and aggregate the resulting answers to form an empirical response distribution (Section 4.4).

4.1 Feature Selection

Feature selection agent identifies a set of contextual variables that are expected to meaningfully influence responses to question q. Rather than selecting from a predefined pool of variables, the framework relies on LLM to reason directly over the question text and determine which contextual attributes are most consequential for shaping response patterns. Given the question q and its answer options, the context agent prompts an LLM to generate a ranked list of potentially relevant demographic or behavioral variables. From the returned list, the top M variables are retained to form:

$$\mathcal{C}_q = \{C_1, \ldots, C_M\}, \tag{1}$$

where M is the number of context variables. In this pipeline, the value of M is not fixed; instead, it is determined dynamically by the LLM based on the inferred complexity and contextual demands of the question.

A key benefit of this design is the transition from a static, predetermined set of contextual variables to a fully adaptive, LLM-driven feature-selection mechanism. By heuristically deciding the context variables, the system can flexibly capture nuanced or previously unrecognized dependencies, "*unknown unknowns*", that would be inaccessible under a fixed-variable regime.

4.2 Distribution Estimation

For each selected contextual variable $C_j \in \mathcal{C}_q$, the context agent constructs an estimated distribution $p(C_j \mid q, g)$ that characterizes plausible values for members of group g when responding to question q. The estimated distributions $p(C_j \mid q, g)$ are not treated as ground truth, but as stochastic priors that induce diversity in downstream generation. Errors in these estimates affect coverage rather than correctness, and their impact is evaluated empirically through fidelity and tail metrics. Normally, there are two types of contextual variables:

Continuous Variables. If C_j is continuous (e.g., age, income), it is modeled using a Gaussian distribution:

$$C_j \sim \mathcal{N}(\mu_j, \sigma_j^2), \tag{2}$$

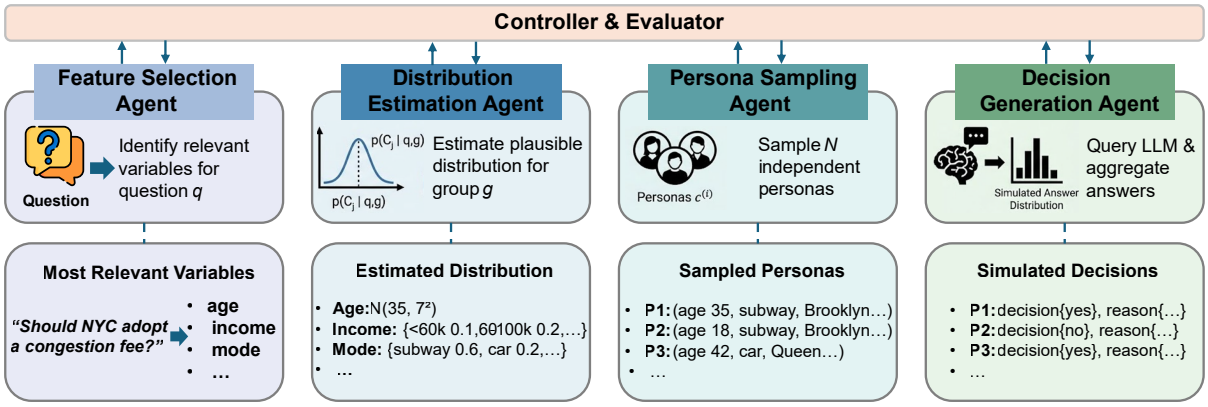

Figure 2: Context modeling framework for LLM-based behavioral simulation.

where the parameters (μ_j, σ_j) are inferred from population metadata, group-level priors, or domain-specific heuristics.

Categorical Variables. If C_j is categorical (e.g., education level, gender), it is represented as a multinomial distribution:

$$\Pr(C_j = r) = \pi_{jr}, \qquad r = 1, \ldots, R_j, \tag{3}$$

where $(\pi_{j1}, \ldots, \pi_{jR_j})$ reflect estimated prevalence across the relevant group and satisfy the normalization constraint $\sum_{r=1}^{R_j} \pi_{jr} = 1$. These probabilities may be obtained from survey statistics, external demographic datasets, or LLM inference when precise information is unavailable.

To simulate full context vectors for each question, it is necessary to specify a joint distribution over all selected variables. To maintain computational efficiency and avoid the combinatorial complexity of modeling high-dimensional dependencies, we simplified the joint distribution as:

$$p(c \mid q, g) = \prod_{j=1}^{M} p(C_j \mid q, g). \tag{4}$$

4.3 Persona Sampling

Given the variable-wise distributions, the context agent generates a collection of N independent persona characteristics intended to represent plausible individuals from group g when responding to question q. Each persona is constructed as a vector of contextual attributes:

$$c^{(i)} = \left(c_1^{(i)}, \ldots, c_M^{(i)}\right), \qquad i = 1, \ldots, N, \tag{5}$$

where each component $c_j^{(i)}$ is sampled independently from its corresponding conditional distribution:

$$c_j^{(i)} \sim p(C_j \mid q, g). \tag{6}$$

This sampling procedure produces a diverse set of synthetic contexts that reflect the variability encoded in the modeled distributions. By drawing multiple independent realizations, the framework captures the heterogeneity of possible respondent profiles rather than relying on a single representation.

4.4 Decision Generation

In decision agent, each sampled persona $c^{(i)}$ is converted into a structured natural-language profile and provided to the decision agent alongside the question q. This representation supplies the LLM ℓ with a detailed description of the contextual scenario under which a response is to be generated. Conditioned on both the question and the sampled context, the LLM produces answers drawn from the induced conditional distribution:

$$a^{(i)} \sim \mathcal{L}_\ell(a \mid q, c^{(i)}). \tag{7}$$

Repeated sampling across independently generated personas yields a collection of responses $\{a^{(i)}\}_{i=1}^{N}$. Aggregating these outputs produces an empirical distribution over answer choices:

$$\hat{p}^{\mathrm{S}}_{m,\ell}(a \mid q, g) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{a^{(i)} = a\}, \tag{8}$$

which serves as an approximation to the simulated response probability under model m and LLM ℓ for the group–question pair (q, g).

This empirical distribution reflects the behavior of the LLM when confronted with a range of plausible contextual samplings and thus approximates the variability inherent in human responses. As N increases, $\hat{p}^{\mathrm{S}}_{m,\ell}(a \mid q, g)$ converges to the underlying conditional distribution induced by the agentic context modeling pipeline.

5 Evaluation Metrics

This study evaluates the quality of LLM-based simulation along four complementary dimensions: *Fidelity*, *Validity*, *Robustness*, and *Consistency*. Let $p^{\mathrm{H}}(\cdot \mid q, g)$ denote the empirical human response distribution for a given question q and demographic group g, and $p^{\mathrm{S}}(\cdot \mid q, g)$ the corresponding distribution produced by the simulation framework.

5.1 Fidelity

Fidelity serves as the primary metric to evaluate how closely the simulated distribution matches the ground truth distribution, capturing the degree to which the context modeling reproduces the empirical response patterns of real participants. This study adopts the Jensen–Shannon divergence (JS) (Fuglede and Topsoe, 2004) as the primary distributional similarity metric due to its symmetry, boundedness, and numerical stability. JS divergence is defined using the Kullback–Leibler divergence (KL), where for two distributions P and Q.

$$\mathrm{KL}(P \,\|\, Q) = \sum_{a} P(a)\, \log \frac{P(a)}{Q(a)} \tag{9}$$

The JS score is computed as:

$$\mathrm{JS} = \tfrac{1}{2}\,\mathrm{KL}\big(p^{\mathrm{H}} \| M\big) + \tfrac{1}{2}\,\mathrm{KL}\big(p^{\mathrm{S}} \| M\big) \tag{10}$$

where $M = \frac{1}{2}\,(p^{\mathrm{H}} + p^{\mathrm{S}})$. Lower values indicate higher fidelity, as the simulated distribution more closely approximates the human distribution.

5.2 Validity

Validity assesses whether the structural relationships present in human data are faithfully preserved by the simulation. This metric examines whether associations between key behavioral or contextual variables are reproduced. For paired variables (X, Y), we measure the deviation between real and simulated correlations:

$$\Delta_{\mathrm{val}} \;=\; \big|\, \mathrm{corr}_{\mathrm{H}}(X, Y) - \mathrm{corr}_{\mathrm{S}}(X, Y) \,\big|, \tag{11}$$

where $\mathrm{corr}_{\mathrm{H}}$ and $\mathrm{corr}_{\mathrm{S}}$ denote Pearson correlations in human and simulated samples, respectively. Smaller values indicate that the simulation preserves behavioral structure more accurately.

5.3 Robustness

Robustness evaluates how sensitive the LLM's simulated responses are to small perturbations, such as prompt rephrasings, minor modifications in persona attributes, or minor changes to LLM parameters. Let $a^{(i)}$ be the answer generated under the original input and $\tilde{a}^{(i)}$ the answer generated under a perturbed version of the same scenario. We quantify robustness via the agreement rate:

$$R \;=\; \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\Big\{ a^{(i)} = \tilde{a}^{(i)} \Big\}. \tag{12}$$

Higher values reflect a more stable simulation, indicating that small, irrelevant variations do not meaningfully affect the output.

5.4 Consistency

Consistency measures whether repeated simulations under identical settings converge to similar aggregate distributions. This metric captures the reproducibility of the simulation pipeline, independent of human data. Let $\hat{P}^{(1)}$ and $\hat{P}^{(2)}$ be two independently generated simulated distributions for the same (q, g) pair. We compute consistency using a normalized Jensen–Shannon similarity:

$$C \;=\; 1 \;-\; \frac{\mathrm{JS}\Big(\hat{P}^{(1)}, \hat{P}^{(2)}\Big)}{\log 2}, \tag{13}$$

This formulation produces a normalized score $C \in [0, 1]$, where higher values reflect stronger consistency and lower values reflect greater divergence.

6 Experiments

6.1 Dataset and Preprocessing

SimBench (Hu et al., 2025) [1] serves as the primary source of survey questions used in this study. SimBench consolidates twenty heterogeneous behavioral datasets (Table 6), ranging from moral dilemmas and personality scales to public-opinion surveys, risk preferences, and numerical estimation, into a unified multiple-choice format. For computational tractability, we randomly sample 100 population-level questions from *SimBenchPop* and 100 group-conditioned question–group pairs from *SimBenchGrouped* per iteration. The data details are in Appendix A.

6.2 Simulation Baselines

Experiments are conducted separately for population-level and group-level settings (Figure 3). At the population level, we compare

[1] All dataset statistics are drawn from the version released at https://huggingface.co/datasets/pitehu/SimBench (downloaded October 2025).

numerical simulation, *direct prompting simulation*, and *context modeling*. At the group level, we compare *numerical simulation*, *group persona prompting*, and *context modeling*. For each question–group pair (q, g), simulated response distributions are evaluated against empirical human data. Detailed baseline descriptions are provided in Appendix D.

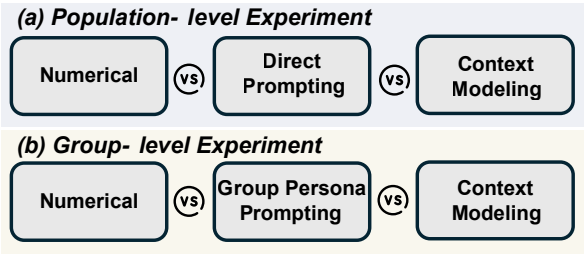

Figure 3: Experimental settings for (a) population-level and (b) group-level experiments.

6.3 Implementation Details

The proposed context modeling pipeline is evaluated across five major LLM families: *ChatGPT*, *Gemini*, *DeepSeek*, *Qwen*, and *Claude*, to assess generalization across architectures and alignment styles. For each model–pipeline combination, responses are simulated for all sampled population-level and group-level questions using $N = 1000$ personas per (q, g) pair. Additional implementation details, including hyperparameters, temperature settings, and prompt templates, are provided in Appendix D.

7 Results and Discussion

7.1 RQ1: Comparative Effectiveness of Context Modeling

Context modeling consistently improves simulation fidelity across all evaluated LLMs, as evidenced by the systematic leftward shift of the JS-distance distributions in Figure 4 and the corresponding reductions in aggregate fidelity reported in Table 1 and Table 2. These gains are observed for every LLM series considered, indicating that the improvement is attributable to the simulation method rather than to model-specific effects. In relative terms, context modeling shows heterogeneous impacts on validity across models, indicating that improvements in distributional fidelity may coexist with modest trade-offs in correlation-based structural accuracy.

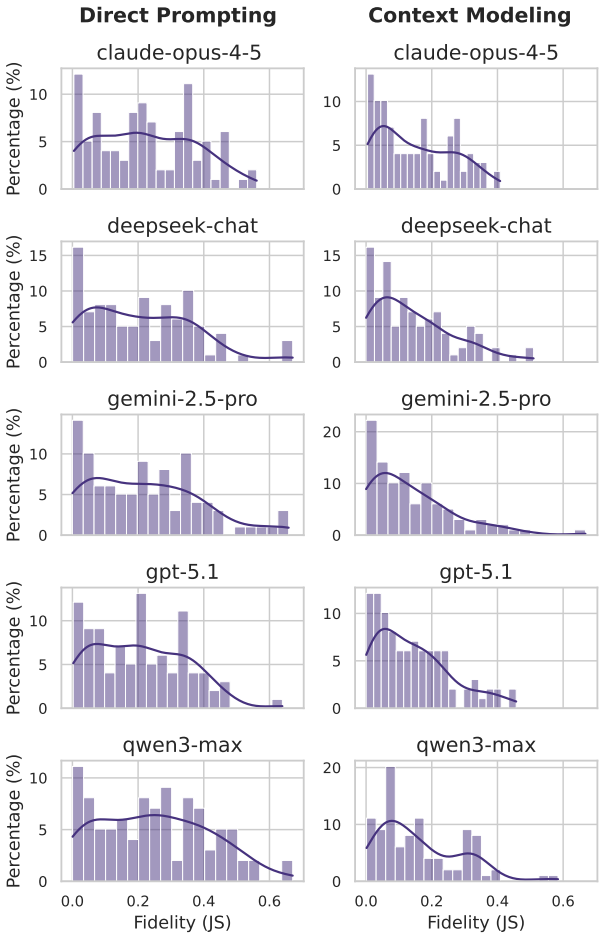

Figure 4: Overall simulation fidelity (JS similarity) for each model and method. Peak to the left is better.

These fidelity improvements, however, are accompanied by a clear tradeoff. As shown in Table 1, robustness and consistency are generally lower under context modeling than under direct prompting. This pattern reflects the design of the context modeling pipeline, which introduces additional stochasticity through context sampling, variable selection, and agentic generation steps. While this stochasticity improves distributional coverage and mitigates mode collapse, thereby enhancing fidelity, it also increases sensitivity to perturbations and run-to-run variability. Notably, the associated reduction in consistency is negligible (below 1% across models), and overall simulation stability remains high.

Table 2: Relative performance improvement by LLMs.

Base Model	ΔFidelity (JS)	ΔValidity (Cor)
Claude-4.5	+33.2%	-15.2%
DeepSeek-V3.2	+31.6%	-21.6%
Gemini-2.5-Pro	**+37.3%**	**+11.4%**
GPT-5.1	+30.8%	+8.0%
Qwen3-Max	+35.9%	-27.6%

Table 1: Overall simulation performance across methods and language models. Fidelity and validity are measured by JS-based distributional distance and correlation-based error, respectively (lower is better). Robustness and consistency are defined in Section 5 (higher is better).

Method	Model	Fidelity (JS)	Validity (Δ_{val})	Robustness(R)	Consistency (C)
Numerical		**0.075 (0.014–0.183)**	0.939	0.600	0.871
Direct Prompting	GPT-5.1	0.204 (0.008–0.454)	0.425	0.847	0.993
	Gemini-2.5-Pro	0.222 (0.003–0.631)	0.427	0.861	0.992
	DeepSeek-V3.2	0.211 (0.008–0.586)	0.513	0.893	**0.997**
	Qwen-3-Max	0.250 (0.009–0.554)	0.385	**0.948**	**0.997**
	Claude-Sonnet-4.5	0.226 (0.010–0.500)	0.506	0.941	**0.997**
Context Modeling	GPT-5.1	0.141 (0.006–0.396)	0.459	0.703	0.987
	Gemini-2.5-Pro	0.139 (0.000–0.437)	0.476	0.666	0.988
	DeepSeek-V3.2	0.144 (0.002–0.421)	0.402	0.689	0.987
	Qwen-3-Max	0.160 (0.007–0.393)	**0.279**	0.664	0.988
	Claude-Sonnet-4.5	0.151 (0.012–0.359)	0.429	0.721	0.989

7.2 RQ2: How Context Modeling Works: Tail Distribution Recovery

One of the major strengths of context modeling is its ability to recover the tail of the human response distribution, including answer options chosen by a minority of respondents. Under direct prompting, LLM-based simulation is governed by next-token probability maximization, which systematically favors the most common or high-probability options in the training distribution. As a result, low-frequency but behaviorally meaningful choices are often assigned near-zero probability, leading to mode collapse in the simulated response distribution. Tail analysis (details in Appendix C.3) demonstrates that context modeling effectively recovers the tail of the human response distribution (Table 3). Across models, it increases the coverage of low-probability but behaviorally meaningful options and reduces the tendency of simulations to collapse onto dominant response modes.

Table 3: Tail comparison of direct prompting (DP) and context modeling (CM).

Metric	DP	CM
Mean tail mass error	0.220	0.196
Mean tail recall	0.243	0.430
Mean tail recall improvement	-	0.187
Mean tail mass error reduction	-	0.024
Mean proportion improved	-	0.337

7.3 RQ3: Performance on Demographic Subgroup Simulation

Context modeling consistently outperforms group persona prompting in representing subgroup populations across base models and demographic dimensions. As shown in Table 5, it achieves systematically lower JS distances, indicating closer alignment with empirical subgroup response distributions. These enhancements come from the simulation framework's explicit modeling of latent, demographically relevant contextual features, which enables more faithful representation of within-group variation across demographic dimensions by capturing heterogeneous preferences, constraints, and situational factors that coarse group-level conditioning cannot express.

Table 4: Average simulation fidelity (JS distance) by demographic variable. Lower is better.

Demographic	CM	GP	Improvement%
Census Region	0.19	0.20	-5.0%
Marital Status	0.12	0.17	23.5%
Political Ideology	0.08	0.30	73.0%
Political Party	0.04	0.07	-860.3%
Religious Attendance	0.00	0.00	0.0%
Age Group	0.11	0.22	50.5%
City Size	0.23	0.31	15.7%
Country	0.13	0.24	31.5%
Discuss Politics	0.11	0.16	33.2%
Domicile	0.12	0.27	35.8%
Education	0.08	0.20	60.4%
ISCED Level	0.18	0.30	39.4%
Employment Status	0.20	0.35	44.3%
Gender	0.11	0.22	47.2%
Highest Education	0.19	0.31	33.3%
Marital Status (Alt.)	0.07	0.20	62.8%
Main Activity	0.14	0.19	28.3%
Religion	0.18	0.26	-66.6%
Religiosity	0.13	0.24	44.8%
Religious Degree	0.02	0.14	69.4%
Subjective Income	0.13	0.19	0.8%
Income Rank	0.13	0.24	49.5%
Urban–Rural	0.04	0.16	70.7%
Work Status	0.22	0.34	29.0%

Note. CM = Context Modeling; GP = Group Persona Prompting.

At the demographic level, context modeling improves fidelity for most subgroup dimensions (Ta-

ble 4), with especially strong gains for attributes characterized by internal diversity. In contrast, dimensions with limited contextual structure or weak behavioral differentiation exhibit negligible or negative changes. These exceptions are limited and do not alter the overall conclusion: context modeling provides a more faithful and flexible representation of subgroup populations than group persona prompting.

Table 5: Subgroup-level simulation fidelity by base model. Fidelity is measured by JS (lower is better).

Model	Context Modeling	Group Prompting	Impr. (%)
Claude-Opus-4.5	**0.103**	0.226	+54.7
DeepSeek-V3.2	**0.134**	0.233	+42.8
Gemini-2.5-Pro	**0.109**	0.236	+53.7
GPT-5.1	**0.133**	0.229	+41.8
Qwen-3-Max	**0.115**	0.226	+49.2

7.4 RQ4: Simulation Performance Across Low- and High-Entropy Questions

Across all evaluated models, context modeling achieves better fidelity on high-entropy (open-ended) questions while preserving strong performance on low-entropy (consensus) questions Figure 5. The lower right-end trend values indicate better performance on high-entropy questions.

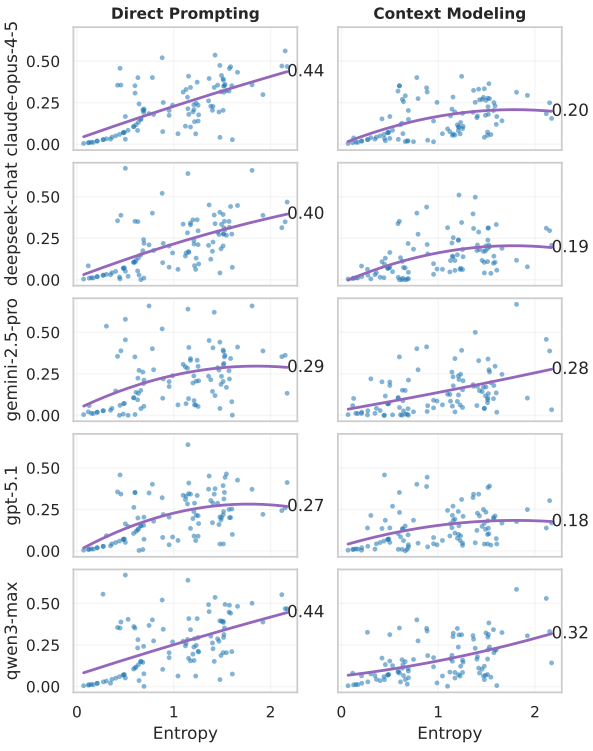

Figure 5: LLM Simulation Fidelity to Question Entropy. The lower trend line is better.

This pattern arises because standard LLM inference is inherently well suited to low-entropy questions, where human responses exhibit strong consensus and limited variability. In such cases, next-token probability maximization naturally aligns with the dominant human choice, yielding high apparent fidelity even under direct prompting. However, high-entropy questions pose a fundamentally different challenge, as they require the simulation to represent a distribution of plausible responses rather than a single mode. Context modeling addresses this limitation by decoupling persona construction from response generation and by introducing structured stochasticity through controlled persona sampling and contextual variation. This design reshapes the model's effective prior over responses, encouraging exploration of multiple plausible decision pathways while maintaining coherence. More details are provided in Appendix C.4

8 Conclusion

This paper demonstrates that agentic context modeling provides a principled and practical foundation for LLM-based behavioral simulation. By separating persona construction from decision generation and injecting structured stochasticity, the approach consistently improves fidelity across question types, demographic groups, and model series. Beyond performance gains, the results highlight a broader insight: realistic behavioral simulation requires explicit representation of heterogeneity rather than reliance on single-shot prompting. Context modeling therefore offers not only a more accurate simulation pipeline, but also a more interpretable and extensible framework for future research on socially informed and policy-relevant AI systems.

Limitations

While context modeling improves simulation fidelity, several limitations remain. First, the independence assumption $p(c \mid q, g) = \prod_j p(C_j \mid q, g)$ may overlook relations among contextual variables; future work could incorporate multivariate or dependency-aware distributions. Second, feature selection relies on LLM heuristics rather than principled criteria, suggesting opportunities for information-theoretic or causal approaches.

Acknowledgments

Omitted for double-blind review.

References

Sameera Abar, Georgios K Theodoropoulos, Pierre Lemarinier, and Gregory MP O'Hare. 2017. Agent based modelling and simulation tools: A review of the state-of-art software. *Computer Science Review*, 24:13–33.

Afrobarometer. 2023. Afrobarometer data, all countries (39), round 9, 2023. http://www.afrobarometer.org. Accessed: March 2025.

Danial Amin, Joni Salminen, Farhan Ahmed, Sonja MH Tervola, Sankalp Sethi, and Bernard J Jansen. 2025. How is generative ai used for persona development?: A systematic review of 52 research articles. *arXiv preprint arXiv:2504.04927*.

Jacy Reese Anthis, Ryan Liu, Sean M Richardson, Austin C. Kozlowski, Bernard Koch, Erik Brynjolfsson, James Evans, and Michael S. Bernstein. 2025. Position: LLM social simulations are a promising research method. In *Forty-second International Conference on Machine Learning Position Paper Track*.

Lora Aroyo, Alex Taylor, Mark Diaz, Christopher Homan, Alicia Parrish, Gregory Serapio-García, Vinodkumar Prabhakaran, and Ding Wang. 2023. Dices dataset: Diversity in conversational ai evaluation for safety. *Advances in Neural Information Processing Systems*, 36:53330–53342.

Tihomir Asparouhov. 2005. Sampling weights in latent variable modeling. *Structural equation modeling*, 12(3):411–434.

Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. The moral machine experiment. *Nature*, 563(7729):59–64.

Edmond Awad, Sohan Dsouza, Azim Shariff, Iyad Rahwan, and Jean-François Bonnefon. 2020. Universals and variations in moral decisions made in 42 countries by 70,000 participants. *Proceedings of the National Academy of Sciences*, 117(5):2332–2337.

Mohammad Beigi, Sijia Wang, Ying Shen, Zihao Lin, Adithya Kulkarni, Jianfeng He, Feng Chen, Ming Jin, Jin-Hee Cho, Dawei Zhou, and 1 others. 2024. Rethinking the uncertainty: a critical review and analysis in the era of large language models. *arXiv preprint arXiv:2410.20199*.

Eric Bigelow and Steven T Piantadosi. 2016. A large dataset of generalization patterns in the number game. *Journal of Open Psychology Data*, 4(1):e4–e4.

Marcel Binz and Eric Schulz. 2023. Turning large language models into cognitive models. *arXiv preprint arXiv:2306.03917*.

Nate Breznau, Eike Mark Rinke, Alexander Wuttke, Hung H. V. Nguyen, Muna Adem, Jule Adriaans, Amalia Alvarez-Benjumea, Henrik K. Andersen, Daniel Auer, Flavio Azevedo, Oke Bahnsen, Dave Balzer, Gerrit Bauer, Paul C. Bauer, Markus Baumann, Sharon Baute, Verena Benoit, Julian Bernauer, Carl Berning, and 147 others. 2022. Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. *Proceedings of the National Academy of Sciences*, 119(44):e2203150119.

Chaoran Chen, Bingsheng Yao, Yanfang Ye, Dakuo Wang, and Toby Jia-Jun Li. 2024. Evaluating the llm agents for simulating humanoid behavior. In *CHI conference proceedingsCHI Conference*. The ACM Conference on Human Factors in Computing Systems-HEAL Workshop (HEAL ....

Alex Clinton, Yiding Chen, Jerry Zhu, and Kirthevasan Kandasamy. 2025. Collaborative mean estimation among heterogeneous strategic agents: Individual rationality, fairness, and truthful contribution. In *Forty-second International Conference on Machine Learning*.

Esin Durmus, Karina Nguyen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, and 1 others. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

Adam Enders, Casey Klofstad, Amanda Diekman, Hugo Drochon, Joel Rogers de Waal, Shane Littrell, Kamal Premaratne, Daniel Verdear, Stefan Wuchty, and Joseph Uscinski. 2024. The sociodemographic correlates of conspiracism. *Scientific reports*, 14(1):14184.

ESS ERIC. 2022. European social survey european research infrastructure. *ESS10-integrated file, edition*, 2.

Bent Fuglede and Flemming Topsoe. 2004. Jensen-shannon divergence and hilbert space embedding. In *International symposium onInformation theory, 2004. ISIT 2004. Proceedings.*, page 31. IEEE.

Ken Goldberg, Theresa Roeder, Dhruv Gupta, and Chris Perkins. 2001. Eigentaste: A constant time collaborative filtering algorithm. *information retrieval*, 4(2):133–151.

ISSP Research Group and 1 others. 2012. International social survey programme: Social inequality iv-issp 2009. *GESIS Data Archive, Cologne. ZA5400 Data file Version*, 3(0).

Onder Gurcan. 2024. Llm-augmented agent-based modelling for social simulations: Challenges and opportunities. *arXiv preprint arXiv:2405.06700*.

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. *arXiv preprint arXiv:2510.17516*.

Minjin Kim, Minju Kim, Hana Kim, Beong-woo Kwak, Soyeon Chun, Hyunseo Kim, SeongKu Kang, Youngjae Yu, Jinyoung Yeo, and Dongha Lee. 2024. Pearl: A review-driven persona-knowledge grounded conversational recommendation dataset. *arXiv preprint arXiv:2403.04460*.

Latinobarómetro. 2023. Latinobarómetro 2023. http://www.latinobarometro.org. Accessed: March 2025.

Dongjun Lee, Juyong Lee, Kyuyoung Kim, Jihoon Tack, Jinwoo Shin, Yee Whye Teh, and Kimin Lee. 2025. Learning to contextualize web pages for enhanced decision making by llm agents. In *International Conference on Representation Learning*, volume 2025, pages 81039–81072.

Jonathan Light, Min Cai, Weiqin Chen, Guanzhi Wang, Xiusi Chen, Wei Cheng, Yisong Yue, and Ziniu Hu. 2025. Strategist: Self-improvement of llm decision making via bi-level tree search. In *International Conference on Representation Learning*, volume 2025, pages 27032–27099.

Gili Lior, Eliya Habba, Shahar Levy, Avi Caciularu, and Gabriel Stanovsky. 2025. Reliableeval: A recipe for stochastic llm evaluation via method of moments. *arXiv preprint arXiv:2505.22169*.

Pingsheng Liu, Zhengjie Huang, Xiechi Zhang, Linlin Wang, Gerard de Melo, Xin Lin, Liang Pang, and Liang He. 2023. A disentangled-attention based framework with persona-aware prompt learning for dialogue generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):13255–13263.

Marlene Lutz, Indira Sen, Georg Ahnert, Elisa Rogers, and Markus Strohmaier. 2025. The prompt makes the person (a): A systematic evaluation of sociodemographic persona prompting for large language models. *arXiv preprint arXiv:2507.16076*.

Niels G Mede, Viktoria Cologna, Sebastian Berger, John Besley, Cameron Brick, Marina Joubert, Edward W Maibach, Sabina Mihelj, Naomi Oreskes, Mike S Schäfer, and 1 others. 2025. Perceptions of science, science communication, and climate change attitudes in 68 countries–the tisp dataset. *Scientific data*, 12(1):114.

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What can we learn from collective human opinions on natural language inference data? *arXiv preprint arXiv:2010.03532*.

Joon Sung Park, Carolyn Q Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S Bernstein. 2024. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*.

Joshua C Peterson, David D Bourgin, Mayank Agrawal, Daniel Reichman, and Thomas L Griffiths. 2021. Using large-scale experiments and machine learning to discover theories of human decision-making. *Science*, 372(6547):1209–1214.

Yilun Qiu, Tianhao Shi, Xiaoyan Zhao, Fengbin Zhu, Yang Zhang, and Fuli Feng. 2025. Latent inter-user difference modeling for llm personalization. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10610–10628.

Joni Salminen, Kathleen Guan, Soon-gyo Jung, Shammur A Chowdhury, and Bernard J Jansen. 2020. A literature review of quantitative persona creation. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–14.

Joni Salminen, Kathleen Wenyun Guan, Soon-Gyo Jung, and Bernard Jansen. 2022. Use cases for design personas: A systematic review and new frontiers. In *Proceedings of the 2022 CHI Conference on human factors in computing systems*, pages 1–21.

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

Gabriel Simmons and Christopher Hare. 2023. Large language models as subpopulation representative models: A review. *arXiv preprint arXiv:2310.17888*.

Camelia Simoiu, Chiraag Sumanth, Alok Mysore, and Sharad Goel. 2019. Studying the "wisdom of crowds" at scale. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 171–179.

Steven Stern. 1997. Simulation-based estimation. *Journal of economic Literature*, 35(4):2006–2039.

Ron Sun. 2018. Cognitive social simulation for policy making. *Policy Insights from the Behavioral and Brain Sciences*, 5(2):240–246.

Anis Tissaoui. 2025. From prompt to persona: a literature review on llms as single cognitive agents. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–17.

SMTI Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*, 6.

Xintao Wang, Heng Wang, Yifei Zhang, Xinfeng Yuan, Rui Xu, Jen tse Huang, Siyu Yuan, Haoran Guo, Jiangjie Chen, Shuchang Zhou, Wei Wang, and Yanghua Xiao. 2025. CoSER: Coordinating LLM-based persona simulation of established roles. In *Forty-second International Conference on Machine Learning*.

A Data Description

Survey questions from *SimBench* comprise a diverse collection of datasets spanning multiple behavioral and decision-making domains, including moral reasoning, social and political attitudes, preference selection, consensus estimation, and socio-economic decision-making. These datasets are drawn from widely used survey instruments and behavioral benchmarks, enabling systematic evaluation of LLM-based behavioral simulation under varied cognitive and uncertainty regimes. Table 6 summarizes the constituent datasets and provides a brief description of their primary behavioral focus.

Questions types. The questions are organized into five high-level categories based on the nature of the underlying task: opinion and attitude, socio-economic and policy, moral dilemma, decision and preference, and consensus and ambiguity. This categorization captures key distinctions in subjectivity, normative structure, and response variability that are central to evaluating behavioral fidelity. Table 7 shows the relative distribution of questions across these domains. *SimBenchPop* consists of broad coverage across all categories, which works as an ideal population-level benchmark spanning both low- and high-entropy questions. In contrast, SimBenchGroup focuses primarily on opinion- and policy-oriented questions, where demographic heterogeneity is well documented and supported by large-scale survey data.

Table 7: Percentage distribution of question categories.

Question Category	SimBenchPop	SimBenchGroup
Opinion	47.1%	78.0%
Socio-economic	7.8%	22.0%
Moral Dilemma	13.7%	0.0%
Decision	15.7%	0.0%
Consensus	15.7%	0.0%

Entropy. To quantify response uncertainty and behavioral diversity, each question is associated with an empirical entropy value computed from the observed human response distribution. Specifically, for a question q with answer options $\mathcal{A}_q$, entropy is defined as

$$H(q) = -\sum_{a \in \mathcal{A}_q} p^{\mathrm{H}}(a \mid q) \log p^{\mathrm{H}}(a \mid q), \quad (14)$$

and for group-conditioned questions in SimBench-Group,

$$H(q, g) = -\sum_{a \in \mathcal{A}_q} p^{\mathrm{H}}(a \mid q, g) \log p^{\mathrm{H}}(a \mid q, g). \quad (15)$$

Lower entropy values indicate strong consensus among respondents, while higher values reflect polarized opinions or intrinsically ambiguous tasks. Figure 6 presents the distribution of entropy values across both benchmarks, highlighting the presence of both consensus-dominated and high-uncertainty questions.

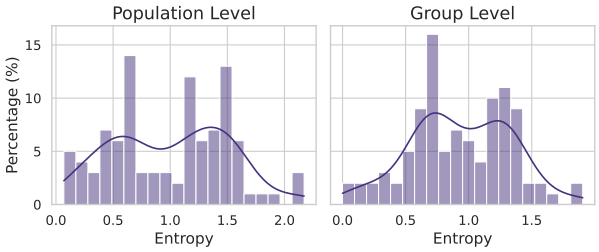

Figure 6: Entropy Distributions for SimBench and Grouped

SimBenchGroup extends the benchmark by explicitly modeling demographic variation in survey responses. Group-conditioned questions are defined over a range of demographic attributes, including country or region, age, gender, education level, employment status, income, political affiliation, religiosity, and urban–rural classification. Each question may involve one or multiple grouping variables, enabling evaluation of whether simulation methods can reproduce both marginal response distributions and systematic cross-group differences. This design supports rigorous testing of demographic fidelity while maintaining comparability across datasets.

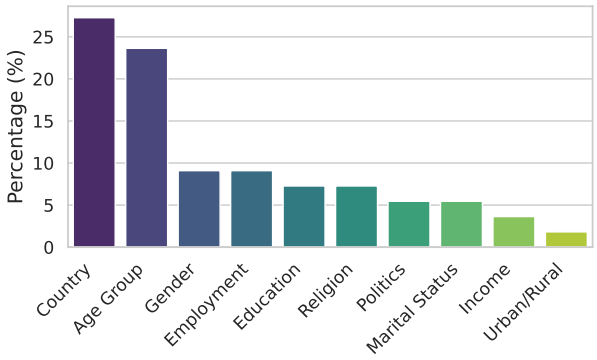

Figure 7: Demographic features tested at group level.

Table 6: Summary of datasets included in SimBench and their primary behavioral domains.

Dataset	Domain	Description
WisdomOfCrowds (Simoiu et al., 2019)	Consensus	Crowd judgments and quantitative estimation tasks.
Jester (Goldberg et al., 2001)	Preference	Humor evaluation and subjective preference choices.
Choices13k (Peterson et al., 2021)	Decision Making	Structured decision tasks across diverse scenarios.
OpinionQA (Santurkar et al., 2023)	Social Attitudes	General attitudes, beliefs, and opinion judgments.
MoralMachineClassic (Awad et al., 2020)	Moral Reasoning	Early moral dilemma scenarios (trolley-type tasks).
MoralMachine (Awad et al., 2018)	Moral Reasoning	Large-scale autonomous-vehicle ethical dilemmas.
ChaosNLI (Nie et al., 2020)	Semantics	Ambiguous NLI judgments adapted to discrete choices.
European Social Survey (ESS) (ERIC, 2022)	Public Opinion	European Social Survey: political and social attitudes.
Afrobarometer(Afrobarometer, 2023)	Public Opinion	Governance and democracy attitudes in Africa.
OSPsycBig5	Personality	Big-Five personality assessment items.
OSPsycMACH	Personality	Machiavellianism and strategic-agency attitudes.
OSPsycMGKT	Mindset	Growth-mindset evaluation items.
OSPsycRWAS	Authoritarianism	Right-wing authoritarianism scale items.
ISSP (Group et al., 2012)	Sociology	Cross-national modules on norms, family, and society.
LatinoBarometro (Latinobarómetro, 2023)	Public Opinion	Political attitudes in Latin American countries.
GlobalOpinionQA (Durmus et al., 2023)	Global Attitudes	International opinions on social and political issues.
DICES (Aroyo et al., 2023)	Econ. Preferences	Incentive-compatible economic choice tasks.
NumberGame (Bigelow and Piantadosi, 2016)	Estimation	Numerical reasoning and quantity estimation.
ConspiracyCorr (Enders et al., 2024)	Beliefs	Conspiracy beliefs and misinformation susceptibility.
TISP (Mede et al., 2025)	Risk / Safety	Safety perceptions and trust in institutions.

B LLMs and Experimental Settings

This study evaluated all simulation pipelines using five representative LLM families that were widely deployed and publicly accessible at the time of experimentation: the ChatGPT, Gemini, DeepSeek, Qwen, and Claude. These models span a range of architectural scales and training paradigms, providing a diverse testbed for assessing the proposed simulation methods. All models were accessed exclusively through their official public APIs.

To ensure comparability across models and simulation pipelines, we adopted a consistent decoding configuration whenever supported by the API. Specifically, the temperature parameter was fixed to 1.0 for all models, corresponding to the default or recommended setting for general-purpose generation. Nucleus sampling (top-p) was left at each model's default value, as default settings are designed to reflect the provider's intended operating regime and are not consistently documented across APIs. For each question (and question–group pair), we generated $N = 1000$ independent samples to approximate response distributions. The same decoding configuration was applied consistently across direct prompting, group persona prompting, and context modeling pipelines to ensure a fair comparison. Table 8 summarizes the language model backends and experimental settings used in our evaluation. Minor model-specific adjustments were made only when required by API constraints, and no task-specific hyperparameter tuning was performed.

Table 8: LLMs and decoding parameters.

Model	Temp.	Top-p	Thinking
GPT-5.1	1.0	0.95	Minimal
GPT-5-Mini	1.0	0.95	Minimal
GPT-5-Nano	1.0	0.95	Minimal
Gemini-2.5-Pro	1.0	0.95	Low
Gemini-2.5-Flash	1.0	0.95	Low
Gemini-2.5-Flash-Lite	1.0	0.95	Low
Qwen3-Max	1.0	0.95	Low
Qwen-Plus	1.0	0.95	Low
Qwen-Flash	1.0	0.95	Low
Claude-3.5-Haiku	1.0	0.95	Low
Claude-Sonnet-4.5	1.0	0.95	Low
Claude-Opus-4.5	1.0	0.95	Low
DeepSeek-V3.2	1.0	0.90	Low

C Additional Results and Discussions

C.1 Overall Performance Ranking

Table 9 reports overall simulation performance across methods and LLMs, ranked by fidelity. The numerical baselines achieve the lowest JS divergence, establishing an empirical upper bound for LLM-based simulation fidelity when ground-truth distributions are directly sampled. This outcome is expected, as numerical simulation draws responses from the observed posterior and therefore minimizes distributional error by construction. However, despite their high fidelity, numerical methods operate purely at the distributional level and do not encode semantic relationships between questions, contextual attributes, and response patterns. As a result, they are limited in their ability to generalize beyond the observed data or to model how responses change under alternative contextual configurations.

Among LLM-based approaches, context modeling consistently outperforms direct prompting across all evaluated models in terms of fidelity, while maintaining strong validity, robustness, and consistency. The improvement is systematic across model series, indicating that gains arise from the simulation framework rather than from specific model architectures or scales.

C.2 Generalization Across Language Model Families and Scales

Context modeling generalizes consistently across language model families and scales, yielding stable fidelity improvements regardless of underlying architecture or parameter size (Table 10). Across both closed- and open-source models, the approach systematically produces lower JS distances than direct prompting, indicating that its effectiveness does not depend on a specific training paradigm or alignment strategy. Importantly, the magnitude of improvement remains comparable across model series, suggesting that the gains arise from the structured simulation pipeline rather than model behavior. Meanwhile, more capable LLMs achieve lower absolute fidelity errors under both prompting strategies, reflecting their stronger semantic understanding and decision coherence. This pattern implies a complementary relationship between model capacity and simulation design: advances in underlying LLMs improve baseline performance, while context modeling consistently reshapes the response distribution to better reflect human heterogeneity.

Table 10: Generalization of context modeling across language model families and scales (lower is better).

Model	Direct Prompt	Context Model.
GPT–5.1	0.222	**0.141**
GPT–5-mini	0.231	**0.153**
GPT–5-nano	0.245	**0.163**
Claude-Opus 4.5	0.226	**0.151**
Claude-Sonnet 4.5	0.219	**0.149**
Claude-Haiku 3.5	0.238	**0.132**
Gemini 2.5 Pro	0.222	**0.139**
Gemini 2.5 Flash	0.228	**0.134**
Qwen3-Max	0.247	**0.160**
Qwen-Plus	0.235	**0.124**
Qwen-Flash	0.229	**0.138**

C.3 Detailed Tail Analysis: Modeling Minority Response Mass

To evaluate how well different simulation methods represent low-probability human responses, we conduct a tail analysis that explicitly focuses on minority answer options. This analysis complements global distributional metrics by isolating regions of the response space where LLMs tend to underrepresent human heterogeneity.

Consider a behavior question q with a finite set of answer options $\mathcal{A}_q = \{a_1, a_2, \ldots, a_K\}$. Let $p^{\mathrm{H}}(a \mid q, g)$ denote the empirical human response distribution for demographic group g. We define the human tail set $\mathcal{T}_{q,g} \subseteq \mathcal{A}_q$ as the set of least frequent answer options whose cumulative probability mass does not exceed a fixed threshold $\tau = 0.2$:

$$\sum_{a \in \mathcal{T}_{q,g}} p^{\mathrm{H}}(a \mid q, g) \leq \tau \tag{16}$$

At least one option is always included in the tail to ensure that the metric is well defined. The total human tail mass serves as the ground truth reference for minority responses:

$$\mathrm{TM}^{\mathrm{H}}_{q,g} = \sum_{a \in \mathcal{T}_{q,g}} p^{\mathrm{H}}(a \mid q, g) \tag{17}$$

For each question group pair, we compare two simulation methods. Direct prompting produces a simulated distribution $p^{\mathrm{B}}(a \mid q, g)$, while context modeling produces $p^{\mathrm{C}}(a \mid q, g)$. The corresponding simulated tail masses are defined as follows:

$$\mathrm{TM}^{\mathrm{B}}_{q,g} = \sum_{a \in \mathcal{T}_{q,g}} p^{\mathrm{B}}(a \mid q, g) \tag{18}$$

$$\mathrm{TM}^{\mathrm{C}}_{q,g} = \sum_{a \in \mathcal{T}_{q,g}} p^{\mathrm{C}}(a \mid q, g) \tag{19}$$

Table 9: Overall simulation performance ranked by fidelity. Fidelity is measured by JS divergence, where lower values indicate higher fidelity. Confidence intervals are reported only for fidelity. Validity, robustness, and consistency are reported as point estimates.

Method	**Model**	**Fidelity** (JS) ↓	**Validity** (Δ_{val})	**Robustness** (R)	**Consistency** (C)
Numerical	Numerical (Full)	0.005 (0.000–0.019)	0.090	0.465	0.973
Numerical	Numerical (Subset)	0.075 (0.014–0.183)	0.061	0.600	0.871
Context Modeling	Qwen Plus	0.124 (0.004–0.390)	0.583	0.649	0.986
Context Modeling	Claude 3.5 Haiku	0.132 (0.003–0.369)	0.537	0.667	0.986
Context Modeling	Gemini 2.5 Flash	0.134 (0.001–0.405)	0.516	0.664	0.985
Context Modeling	Qwen Flash	0.138 (0.001–0.434)	0.606	0.641	0.984
Context Modeling	Gemini 2.5 Pro	0.139 (0.000–0.437)	0.524	0.666	0.988
Context Modeling	GPT-5.1	0.141 (0.006–0.396)	0.541	0.703	0.987
Context Modeling	DeepSeek Chat	0.144 (0.002–0.421)	0.598	0.689	0.987
Context Modeling	Claude Sonnet 4.5	0.149 (0.007–0.378)	0.560	0.671	0.987
Context Modeling	Claude Opus 4.5	0.151 (0.012–0.359)	0.559	0.681	0.985
Context Modeling	GPT-5 Mini	0.153 (0.005–0.470)	0.698	0.681	0.985
Context Modeling	Qwen3 Max	0.160 (0.007–0.393)	0.721	0.664	0.988
Context Modeling	GPT-5 Nano	0.163 (0.002–0.516)	0.540	0.719	0.986
Direct Prompting	GPT-5.1	0.204 (0.008–0.454)	0.575	0.847	0.993
Direct Prompting	DeepSeek Chat	0.211 (0.008–0.586)	0.487	0.893	0.997
Direct Prompting	Gemini 2.5 Pro	0.222 (0.003–0.631)	0.573	0.861	0.992
Direct Prompting	Claude Opus 4.5	0.226 (0.010–0.500)	0.494	0.941	0.997
Direct Prompting	Qwen3 Max	0.250 (0.009–0.554)	0.615	0.948	0.997

Tail mass fidelity is measured using absolute error relative to the human distribution. Lower values indicate closer alignment with the human minority mass:

$$\text{Err}^{\text{B}}_{q,g} = \left|\text{TM}^{\text{B}}_{q,g} - \text{TM}^{\text{H}}_{q,g}\right| \tag{20}$$

$$\text{Err}^{\text{C}}_{q,g} = \left|\text{TM}^{\text{C}}_{q,g} - \text{TM}^{\text{H}}_{q,g}\right| \tag{21}$$

To quantify improvement from context modeling, we define tail error reduction as the difference between baseline and context errors:

$$\Delta\text{Err}_{q,g} = \text{Err}^{\text{B}}_{q,g} - \text{Err}^{\text{C}}_{q,g} \tag{22}$$

In addition to aggregate probability mass, we evaluate whether individual minority options receive non-negligible probability. Let $\varepsilon = 0.01$ denote a small probability threshold. Tail recall measures the fraction of tail options assigned probability at least ε:

$$\text{Recall}^{\text{M}}_{q,g} = \frac{1}{|\mathcal{T}_{q,g}|} \sum_{a \in \mathcal{T}_{q,g}} \mathbf{1}\left[p^{\text{M}}(a \mid q, g) \geq \varepsilon\right] \tag{23}$$

Here $\text{M} \in \{\text{B}, \text{C}\}$ indexes the simulation method. Tail recall improvement from context modeling is defined as follows:

$$\Delta\text{Recall}_{q,g} = \text{Recall}^{\text{C}}_{q,g} - \text{Recall}^{\text{B}}_{q,g} \tag{24}$$

All tail metrics are computed for each valid model question group triple where both simulation methods are available. Results are aggregated by model, reporting mean values and 95% confidence intervals across all evaluated pairs. We also report the fraction of pairs for which tail error reduction is positive, providing an interpretable measure of how consistently context modeling improves minority response fidelity.

Table 11: Tail analysis metrics aggregated by model. Lower values indicate better performance for tail mass error, while higher values indicate better performance for tail recall and improvement metrics.

Metric	**DP**	**CM**
Mean tail mass error	Err^{B}	Err^{C}
Mean tail recall	Recall^{B}	Recall^{C}
Tail mass error reduction	ΔErr	
Tail recall improvement	ΔRecall	
Fraction of improved pairs	$\Pr(\Delta\text{Err} > 0)$	

The result provides clear evidence that context modeling improves LLM-based behavioral simulation primarily by strengthening the representation of minority response options rather than merely rescaling aggregate probabilities. As shown in Table 12, context modeling substantially increases tail recall for all evaluated models, with improvements ranging from 9.2% to 26.5%. This indicates that, under direct prompting, LLMs frequently collapse low-probability human responses to near-zero likelihood, whereas context modeling systematically

Table 12: Tail analysis summary across language models. Tail mass error measures absolute deviation from the human minority mass, where lower values indicate better fidelity. Tail recall measures the fraction of minority options assigned non-negligible probability, where higher values indicate better coverage. Percentage Improved denotes the proportion of question–group pairs for which context modeling reduces tail mass error relative to direct prompting.

Metric	Claude Opus 4.5	DeepSeek Chat	Gemini 2.5 Pro	GPT-5.1	Qwen3 Max
Direct Prompting tail mass error	0.203	0.192	0.233	0.215	0.256
Context Modeling tail mass error	0.186	0.193	0.176	0.198	0.228
Tail mass error reduction	0.017	−0.000	0.057	0.017	0.027
Direct Prompting tail recall	16.8%	20.6%	29.9%	31.2%	23.1%
Context Modeling tail recall	38.3%	45.0%	41.9%	40.4%	49.6%
Tail recall improvement	21.5%	24.4%	12.0%	9.2%	26.5%
Percentage Improved	31.3%	34.3%	36.4%	33.3%	33.3%

reallocates probability mass across a broader set of minority options. Again, these recall gains are consistent across model families, suggesting that the effect arises from the simulation pipeline rather than model-specific characteristics. Improvements in tail mass fidelity are more moderate but remain positive for most models, with tail mass error reductions observed in approximately one-third of question–group pairs. This pattern implies that context modeling first mitigates mode collapse by preserving minority options, and only secondarily improves the precise calibration of their aggregate probability mass.

C.4 Additional Results on RQ4

High-entropy questions represent cases where human responses exhibit substantial disagreement and behavioral heterogeneity, making them particularly challenging for LLM-based simulation. Focusing on the top 30% of questions by entropy, context modeling achieves a substantial improvement in simulation fidelity relative to direct prompting. Specifically, the mean JS divergence decreases from 0.127 under direct prompting to 0.086 under context modeling, corresponding to an improvement of approximately 32%. This reduction indicates that context modeling is markedly more effective at capturing the dispersed response distributions characteristic of high-entropy settings. In contrast, direct prompting exhibits significantly higher divergence, reflecting its tendency to overconcentrate probability mass on dominant response modes. As expected, numerical simulation achieves the lowest divergence, serving as an empirical upper bound, but lacks the ability to model contextual dependencies or generalize beyond observed posteriors.

Table 13: Simulation fidelity on high-entropy questions (top 30% by entropy). Lower JS indicates higher fidelity.

Method	Mean JS	Median JS
Numerical	0.002	0.002
Context Modeling	0.086	0.076
Direct Prompting	0.127	0.173

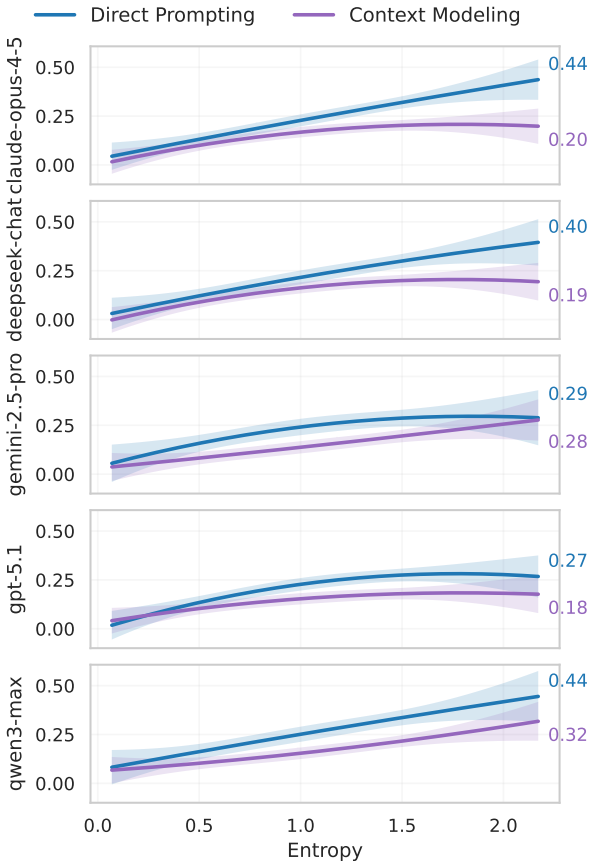

Figure 8: Performance comparison between direct prompting and context modeling via entropy.

Context modeling is particularly effective for high-entropy questions because such settings re-

quire the simulation to represent genuine behavioral diversity rather than converge to a single dominant response. Under direct prompting, LLM inference is driven by next-token probability maximization, which favors the most salient or frequent options and leads to overconfident, low-entropy outputs even when human responses are highly dispersed. Context modeling mitigates this limitation by explicitly structuring contextual attributes and sampling across diverse personas, thereby reshaping the effective response prior and encouraging exploration of multiple plausible decision pathways. This structured stochasticity allows the model to allocate probability mass across competing responses in a controlled manner, aligning simulated distributions more closely with the heterogeneous patterns observed in human data.

C.5 Additional Discussion

Simulation fidelity depends more on where randomness is introduced than on how much randomness is used. Increasing stochasticity solely at the decoding stage, such as by raising temperature or adjusting sampling parameters, often amplifies noise without addressing systematic biases in LLM-based simulation. This approach can degrade performance on low-entropy questions while failing to recover meaningful diversity on high-entropy ones. In contrast, context modeling introduces randomness at the level of persona construction and contextual conditioning, where variability corresponds to interpretable dimensions of human heterogeneity. By placing stochasticity upstream in the simulation pipeline, context modeling enables controlled exploration of diverse response tendencies, leading to improved distributional fidelity without sacrificing robustness or consistency.

In addition, larger models do not automatically solve the challenges of behavioral simulation, and structural design remains essential. Despite their increased capacity and stronger language understanding, larger LLMs still exhibit majority bias and mode collapse under direct prompting, particularly in settings that require modeling heterogeneous human behavior. The consistent improvements achieved by context modeling across model sizes and families indicate that simulation performance is driven primarily by how contextual information is represented and utilized rather than by scale alone. These findings underscore that effective behavioral simulation requires explicit structural mechanisms to capture context and heterogeneity, even when using state-of-the-art large language models.

D Simulation Pipelines and Prompt Template

D.1 Numerical Simulation

The numerical simulation pipeline serves as a non-LLM baseline that generates synthetic responses using only observed survey statistics. The core idea is to treat each question as a discrete choice problem and simulate responses by sampling from the empirical human response distribution estimated from ground-truth data. For a question q with answer set $\mathcal{A}_q = \{a_1, \ldots, a_K\}$, let $n_q(a)$ denote the number of respondents who selected option a, and let $N_q = \sum_{a \in \mathcal{A}_q} n_q(a)$ be the total sample size. The empirical response distribution is then computed as

$$\hat{p}^{\mathrm{H}}(a \mid q) = \frac{n_q(a)}{N_q}, a \in \mathcal{A}_q. \tag{25}$$

To simulate S independent synthetic responses for question q, we draw

$$\tilde{y}^{(s)} \sim \mathrm{Categorical}\big(\hat{p}^{\mathrm{H}}(\cdot \mid q)\big)\,, s = 1, \ldots, S, \tag{26}$$

and aggregate the simulated outcomes into a simulated distribution

$$\hat{p}^{\mathrm{Num}}(a \mid q) = \frac{1}{S} \sum_{s=1}^{S} \mathbb{I}\Big[\tilde{y}^{(s)} = a\Big]\,, a \in \mathcal{A}_q, \tag{27}$$

where $\mathbb{I}[\cdot]$ is the indicator function. When demographic groups are available, the same procedure applies conditionally: for group g, we compute $\hat{p}^{\mathrm{H}}(a \mid q, g) = n_{q,g}(a)/N_{q,g}$ and sample from $\mathrm{Categorical}(\hat{p}^{\mathrm{H}}(\cdot \mid q, g))$. This pipeline isolates the effect of sampling variability and provides a principled reference point for evaluating LLM-based simulation methods, since it reflects the best achievable performance under a purely distributional resampling assumption.

Example. Consider a survey question:

Example Survey Question

Question: There are some things that can be done to reduce energy use, such as switching off appliances that are not being used, walking for short journeys, or only using the heating or air conditioning when really needed. In your daily life, how often do you do things to reduce your energy use?

Answer options:
(A) Never (B) Hardly ever (C) Sometimes
(D) Often (E) Very often (F) Always
(G) Cannot reduce energy use

Suppose a total of $N_q = 1000$ survey responses are observed for the energy-use question, distributed across the seven answer options. Using Eq. (25), these counts are normalized to obtain an empirical response distribution $\hat{p}^{\mathrm{H}}(\cdot \mid q)$, with the highest mass placed on mid-to-high frequency responses (e.g., "Often" and "Very often").

To generate synthetic data, we draw $S = 1,000$ responses from a categorical distribution parameterized by $\hat{p}^{\mathrm{H}}(\cdot \mid q)$ (Eq. (26)). Aggregating these samples yields a simulated distribution $\hat{p}^{\mathrm{Num}}(\cdot \mid q)$ that closely matches the empirical distribution, with only small deviations due to finite-sample Monte Carlo variability. As S increases, $\hat{p}^{\mathrm{Num}}(\cdot \mid q)$ converges to $\hat{p}^{\mathrm{H}}(\cdot \mid q)$, confirming that numerical simulation preserves the original response proportions by construction.

D.2 Direct Prompting Simulation

Direct prompting simulation serves as a baseline LLM-based approach in which the model is queried to produce a single survey response conditioned only on the question text and, when applicable, minimal demographic descriptors. Unlike agentic or context-modeling methods, this approach does not explicitly model uncertainty, population distributions, or iterative reasoning. Instead, it relies on the LLM's implicit knowledge and internal priors to select the most plausible answer. For each question, the model is prompted to choose exactly one option from the predefined answer set, and repeated queries are used to approximate a response distribution. This method reflects a common practice in prior work. Figure 9 shows the prompt for direct prompting of the example question.

Direct Prompting Template

Question: There are some things that can be done to reduce energy use, such as switching off appliances that are not being used, walking for short journeys, or only using the heating or air conditioning when really needed. In your daily life, how often do you do things to reduce your energy use?
Answer options: (A) Never (B) Hardly ever
(C) Sometimes (D) Often (E) Very often
(F) Always (G) Cannot reduce energy use Given the characteristics above, pick the single option this person would choose. Valid answers (letters only): A, B, C, D, E, F, G. If unsure, choose the most plausible option. Your response MUST be exactly one of the valid letters. Return only that single letter (e.g., A).

Figure 9: Prompt template for direct prompting.

D.3 Group Persona Prompting Simulation

Group persona prompting extends direct prompting by explicitly conditioning the LLM on a predefined demographic profile intended to represent a specific population group. Instead of relying solely on the question text, the model is provided with group-level characteristics such as country, age group, gender, education, or other socio-demographic attributes. The LLM is instructed to answer the survey question from the perspective of a typical individual belonging to the specified group. While this approach introduces demographic context, it still operates in a single-step manner and does not explicitly model response uncertainty or population-level distributions. Repeated queries are therefore required to approximate a response distribution. Group persona prompting is widely used in applied LLM studies and serves as a natural intermediate baseline between direct prompting and more structured context-modeling pipelines. Figure 10 illustrates the prompt used for group persona prompting with the example question.

Group Persona Prompting Template

Persona Description:
You are asked to simulate the response of a randomly selected individual from the following demographic group:
Occupation: Student
Question:
There are some things that can be done to reduce energy use, such as switching off appliances that are not being used, walking for short journeys, or only using the heating or air conditioning when really needed. In your daily life, how often do you do things to reduce your energy use?

Answer options:
(A) Never (B) Hardly ever (C) Sometimes
(D) Often (E) Very often (F) Always
(G) Cannot reduce energy use

First, randomly sample a plausible individual from the demographic group described above. Then, from that individual's perspective, pick the single option this person would choose.
Valid answers (letters only): A, B, C, D, E, F, G. If unsure, choose the most plausible option. Your response MUST be exactly one of the valid letters. Return only that single letter (e.g., A).

Figure 10: Prompt template for group persona prompting with prompt-based random individual sampling.

D.4 Context Modeling

Context modeling adopts a structured, multi-agent simulation pipeline to generate survey responses that explicitly account for contextual relevance, demographic heterogeneity, and response uncertainty. Unlike direct prompting or group persona prompting, which rely on a single-step decision conditioned on a fixed description, context modeling decomposes the simulation process into four coordinated agents: *feature selection*, *distribution estimation*, *sampling*, and *decision*. Each agent performs a well-defined role, enabling the simulation to separate contextual reasoning from stochastic response generation.

The pipeline begins with a *feature selection agent*, whose role is to identify the subset of respondent characteristics that are most relevant to the given survey question. Conditioning on all available attributes can introduce noise and spurious correlations; therefore, this agent selects only salient demographic and attitudinal variables based on the semantic content of the question. For the energy-use frequency question, features such as country, age, education level, income, and environmental concern are typically prioritized.

Feature Selection Agent Prompt

You are an API that returns **ONLY JSON**. Do not include any text outside the JSON output.
Your task is to identify the most relevant demographic and psychographic characteristics that are likely to influence how respondents answer the survey question below. You are *not* given a predefined feature pool and should freely select the most salient and interpretable characteristics.
Output format:
Return a JSON array containing the top k features, ranked from most to least relevant. Each element in the array *must* follow exactly the schema below:

```
{"feature": "<name>", "rank": <integer>}
```

Rules:
(1) Use *exactly* the keys shown above.
(2) Ranks must be consecutive integers starting from 1.
(3) Feature names should be concise, descriptive, and commonly used in survey research.
(4) Do *not* include explanations, comments, or additional fields.
(5) Return *only* the JSON array.
Survey Question:
There are some things that can be done to reduce energy use, such as switching off appliances that are not being used, walking for short journeys, or only using the heating or air conditioning when really needed. In your daily life, how often do you do things to reduce your energy use?

Given the selected features, the distribution estimation agent constructs a conditional population-level distribution over respondent characteristics for the specified question–group pair. This distri-

bution represents heterogeneity within the group and serves as the basis for sampling individual instances.

The sampling mechanism provides several key benefits that are not attainable through direct or single-persona prompting. Most importantly, it enables the simulation to capture individuals in the tails of the population distribution, respondents whose behaviors, constraints, or preferences deviate from a notional "average" or fully rational agent. Direct prompting typically elicits responses that reflect a coherent, norm-following, and internally consistent persona, which tends to overrepresent modal or socially desirable behavior and underrepresent marginal, inconsistent, or constrained decision patterns observed in real survey data. By contrast, distribution-based sampling explicitly generates respondents across the full support of relevant characteristics.

Distribution Estimation Agent Prompt

You are an API that returns **ONLY JSON**. Do not include any text outside the JSON output.
Your task is to estimate the population-level distribution of the most relevant demographic and psychographic characteristics for respondents answering the survey question below.

Selected characteristics:
country, age, household_income_level, education_level, environmental_concern
Output format:
Return a JSON array containing the top k features. Each element in the array *must* follow exactly one of the two schemas below:

Continuous feature

```
{"feature": "<name>", "type": "continuous",
"mean": <float>, "std": <float>,
"bounds": [<min>, <max>]}
```

Categorical feature

```
{"feature": "<name>", "type": "categorical",
"distribution":
{"label1": <prob>, "label2": <prob>}}
```

Rules:
(1)Use *exactly* the keys shown above. (2)Probabilities must sum to approximately 1.0. (3) Bounds must be a two-element array `[min, max]`. (4) Do *not* include explanations, comments, or additional fields.(5)Return *only* the JSON array.
Survey Question:
There are some things that can be done to reduce energy use, such as switching off appliances that are not being used, walking for short journeys, or only using the heating or air conditioning when really needed. In your daily life, how often do you do things to reduce your energy use?

Distribution Estimation Agent Output

```
{
  "cntry": {
    "type": "categorical",
    "distribution": { "Italy": 1.0 }
  },
  "age": {
    "type": "continuous",
    "mean": 35.0,
    "std": 10.0,
    "bounds": [18.0, 80.0]
  },
  "household_income_level": {
    "type": "categorical",
    "distribution": {
      "low": 0.4,
      "middle": 0.4,
      "high": 0.2
    }
  },
  "education_level": {
    "type": "categorical",
    "distribution": {
      "high_school_or_less": 0.5,
      "some_college": 0.3,
      "bachelor_or_higher": 0.2
    }
  },
  "environmental_concern": {
    "type": "continuous",
    "mean": 3.0,
    "std": 1.0,
    "bounds": [1.0, 5.0]
  }
}
```

The *sampling agent* is responsible for instantiating a concrete synthetic individual by stochastically sampling from the conditional feature distributions estimated in the previous stage. This step operationalizes within-group heterogeneity by translating population-level distributions into individual-level realizations. In implementation, standard pseudo-random sampling routines (e.g., `np.random`) are used, with controlled random seeds to ensure reproducibility.

Let $\mathcal{C}_q$ denote the set of selected contextual features for question q. For a categorical feature $c \in \mathcal{C}_q$ with support $\{v_1, \ldots, v_K\}$ and associated probabilities $\{\pi_1, \ldots, \pi_K\}$, the sampling agent draws

$$\tilde{c} \sim \text{Categorical}(\pi_1, \ldots, \pi_K). \tag{28}$$

For a continuous feature c parameterized by mean μ_c and standard deviation σ_c, a sample is drawn according to

$$\tilde{c} \sim \mathcal{N}(\mu_c, \sigma_c^2), \tag{29}$$

and subsequently truncated to satisfy the prescribed bounds. The resulting set of sampled values $\tilde{\mathbf{c}} =$

$\{\tilde{c} : c \in \mathcal{C}_q\}$ defines a plausible synthetic individual consistent with the estimated conditional distribution. Repeating this procedure across multiple runs ensures that the simulation captures natural within-group variation rather than collapsing to a deterministic or modal outcome.

Finally, the *decision agent* outputs the sampled response in a strictly constrained format suitable for evaluation. This agent enforces compliance with the predefined answer set and suppresses any additional explanation, ensuring consistency across simulation methods.

Decision Agent Prompt

You are a simulated survey participant with the following characteristics:
- Country: Italy
- Age: 39.64
- Household income level: middle
- Education level: bachelor_or_higher
- Environmental concern: 1.98

Question: There are some things that can be done to reduce energy use, such as switching off appliances that are not being used, walking for short journeys, or only using the heating or air conditioning when really needed. In your daily life, how often do you do things to reduce your energy use?
Options: (A) Never (B) Hardly ever (C) Sometimes (D) Often (E) Very often (F) Always (G) Cannot reduce energy use
Pick the single option this person would choose. Valid answers (letters only): A, B, C, D, E, F, G. If unsure, choose the most plausible option. Your response MUST be exactly one of the valid letters. Return only that single letter. No words, no punctuation, no explanations.

By repeating this four-stage process, context modeling produces a simulated response distribution that reflects both systematic contextual effects and individual-level variability. This design enables more faithful behavioral simulation than single-step prompting, while remaining interpretable and modular.