Data-Driven Personas for Survey Simulation: Insights into Simulation Alignment Across Data-Access Regimes

Dongryeol Lee*
Seoul National University
drl123@snu.ac.kr

Weronika Łajewska
Amazon
lajewska@amazon.com

Leonardo Perelli
Amazon
lperelli@amazon.com

Saab Mansour
Amazon
saabm@amazon.com

Abstract

Large language models (LLMs) offer new opportunities for public opinion research by enabling early prediction of survey responses, potentially reducing the cost and time of traditional surveys. However, many existing steering approaches rely on target-domain human data for fine-tuning or prompting that is costly to collect and raises privacy concerns. In this paper, we study demographic group-level survey simulation, where personas induced from heterogeneous, anonymized public behavioral data condition agents that simulate responses of individuals from specific demographic groups. We examine whether representative personas can be induced from diverse sources and analyze how the domain, scale, and granularity of the source data affect survey simulation alignment. We find that personas induced from out-of-domain sources rarely outperform simulations conditioned only on basic demographic information, largely due to population mismatch. However, when personas are accurately assigned to the target demographic groups, alignment improves substantially. Finally, personas induced from target-domain survey data generalize better as more survey question history becomes available, suggesting that richer behavioral evidence enables more stable persona trait inference that transfers to better unseen questions simulation alignment.

1 Introduction

Surveys are an essential tool for understanding public opinion and informing decisions in public policy, industry, and social science research. Because attitudes and preferences evolve over time and vary across demographic groups, organizations often rely on repeated and large-scale surveys to capture these differences. However, collecting survey data is costly, time-consuming, and logistically demanding. Recent advances in large language models (LLMs) have enabled their use for *survey simulation*, where models are prompted to predict the opinions of different demographic groups (Wang et al., 2025; Gao et al., 2024; Manning et al., 2024; Bail, 2024). This approach offers a scalable alternative to traditional user studies while enabling faster exploration of public opinion dynamics (Argyle et al., 2023; Aher et al., 2023; Hewitt et al., 2024).

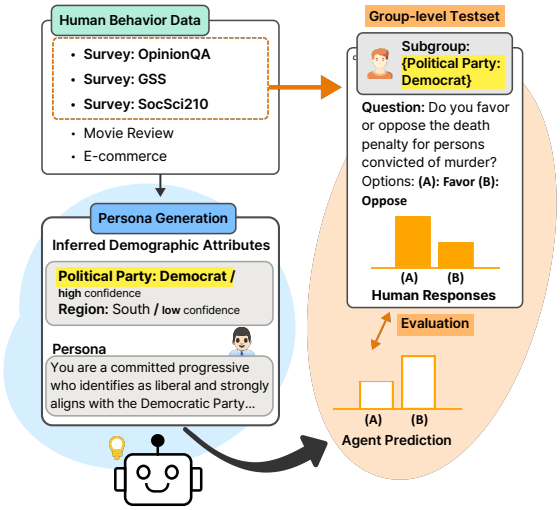

Figure 1: Data driven persona induction and survey simulation: In our framework, we generate personas from behavioral data across different domains and predict responses for individuals in target demographic groups.

Prior work on survey simulation has primarily explored prompt engineering (Santurkar et al., 2023; Kim and Lee, 2023), persona-based conditioning (Park et al., 2024a), and fine-tuning (Cao et al., 2025; Kolluri et al., 2025; Suh et al., 2025) to improve the alignment between simulated and human responses. However, much of this work relies on target-domain human data for learning and control—for example, by using interviews or other

* Work done as an intern at Amazon.

individual-level records in prompts (Park et al., 2024a), or fine-tuning models on data drawn from the same source as the evaluation set (Cao et al., 2025; Kolluri et al., 2025; Suh et al., 2025). Collecting such data is resource-intensive and often constrained by privacy regulations, consent requirements, and governance frameworks, while also providing limited coverage of underrepresented subpopulations.

In this work, we investigate the impact of synthetic personas induced from heterogeneous sources of human behavioral data on steering LLM-based survey simulations. [1] Our goal is to understand when such personas, defined as natural-language profiles distilled from behavioral traces, improve the alignment of survey simulation. This question is important in practice because practitioners often have uneven access to persona source data. In some settings, only external behavioral data are accessible (**Low access / out-of-domain**); in others, additional information such as ground-truth (GT) demographics of participants may be accessible (**Mid access / out-of-domain with GT demographics**); and in the most favorable cases, part of the target survey data itself may be available (**High access / in-domain**). To reflect these realistic constraints, we frame persona-conditioned survey simulation in terms of **data-access regimes** and study how the effectiveness of personas changes as stronger forms of supervision become available.

Accordingly, we treat persona induction from behavioral data and downstream survey simulation as separate tasks. We first induce personas from different sources of behavioral data, then use them to condition individual agents, and evaluate the resulting aggregated predictions against human response distributions at the demographic-group level (e.g., the Democrat group distribution in Figure 1). [2]

Across survey benchmarks, we find that synthetic personas are not uniformly beneficial for simulation—their effectiveness depends strongly on the available data and the degree of source–target alignment. Nevertheless, the value of persona-conditioned simulation should not be judged solely by aggregate alignment. Persona-based simulation can also offer an important advantage in interpretability: each simulated response is linked to a specific persona induced from behavioral evidence, and aggregating over multiple personas allows the simulator to represent within-group heterogeneity rather than collapsing a group into a single demographic descriptor. We therefore view personas not only as a potential means of improving aggregate alignment, but also as a way to make survey simulation more behaviorally grounded, more interpretable, and more sensitive to within-group variation. More broadly, our results provide practical guidance for deploying survey simulation under realistic data constraints, showing that effective simulation depends not only on the richness of persona information, but also on the type of behavioral data, accuracy of persona to group assignment, and source–target population mismatch (i.e., the degree to which the source behavioral population resembles the target survey respondents).

Our contributions are summarized as follows:

- We show that personas induced from out-of-domain data do not generalize well to survey simulation, largely because demographic attribute misassignment causes personas to be mapped to inappropriate target groups.
- We show that personas induced from in-domain data generalize better to held-out questions as more history is provided. They also yield the strongest and most stable alignment across different sizes of demographic groups, with especially large gains for smaller, underrepresented groups, and perform best across all demographic attributes.

2 Related Work

Survey simulation. Recent work explores LLMs as a lower-cost alternative to large-scale human data collection, demonstrating their ability to reproduce some group-level response patterns in surveys and behavioral, economic, and political settings (Argyle et al., 2023; Aher et al., 2023; Horton, 2023; Hewitt et al., 2024). However, subsequent studies raised concerns about validity, robustness, and representational bias (Dominguez-Olmedo et al., 2024; Bisbee et al., 2024). Subsequent efforts proposed prompting and elicitation methods, including sociodemographic prompting,

[1] All experiments use publicly available datasets with anonymized behavioral traces, and no personally identifiable information was used or collected in this research. To support reproducibility, we release the code at https://github.com/amazon-science/data-driven-personas, but not raw user-level traces, record-linked personas, or intermediate artifacts that could increase re-identification risk.

[2] We do not aim to replicate individual behavior, but to statistically approximate group-level responses.

persona-based conditioning, and elicitation of response distributions (Santurkar et al., 2023; Moon et al., 2024; Park et al., 2024a; Manning et al., 2024; Kwok et al., 2024; Sun et al., 2024). Yet capturing within-group heterogeneity and avoiding stereotypical predictions remain open challenges (Park et al., 2024b; Kapania et al., 2025; Cheng et al., 2023), and fine-tuning approaches rely on survey or experimental response data for adaptation (Cao et al., 2025; Kolluri et al., 2025; Suh et al., 2025). In contrast, we focus on simulation of *group-level response distributions* and examine how conditioning signals transfer across domains and data-access regimes.

Synthetic personas. Synthetic persona generation constructs artificial user profiles to support downstream tasks such as political and economic simulation (Tseng et al., 2024; Chen et al., 2024; Horton, 2023). Earlier approaches often relied on survey or census data (Argyle et al., 2023; Chang et al., 2025), demographic profiles (Hwang et al., 2023; Santurkar et al., 2023), or political identities (Simmons, 2023) to steer model behavior. With LLMs, recent work has moved toward descriptive personas derived from richer sources, including web text (Ge et al., 2024), interview transcripts (Park et al., 2024a), personal narratives (Moon et al., 2024), and shopping histories (Mansour et al., 2025), with behavioral-data-grounded personas also supporting A/B test simulation (Benomar et al., 2026). Other work further explores automatically refined personalization: few-shot personalization (Kim and Yang, 2025) and collaborative data refinement (Sun et al., 2025). These methods provide richer conditioning signals, but approaches grounded in individual records may depend on costly or sensitive target-domain data, and demographic prompting can yield caricatured simulations (Cheng et al., 2023). In contrast, we focus on *data-driven personas induced from external behavioral sources* to steer *distribution-level* survey simulation, enabling out-of-domain deployment without individual-level target-survey records.

3 Survey Simulation under Different Data-access Regimes

In this work, we study *agentic survey simulation*, in which a pool of LLM-based agents predict how humans would respond to survey questions (Argyle et al., 2023; Wang et al., 2025). Our central question is whether—and under what conditions—

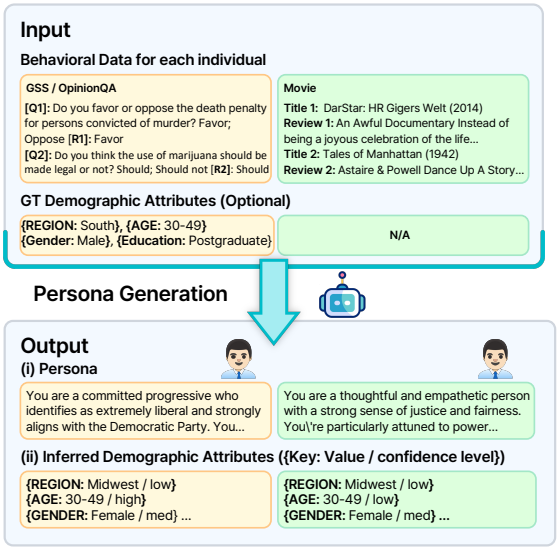

Figure 2: Persona generation examples across domains.

synthetic personas, defined as natural-language profiles inferred from behavioral data, improve simulation alignment for survey responses. In our setting, agents are conditioned on synthetic personas, while evaluation is performed on aggregated response distributions for each demographic group (Suh et al., 2025; Cao et al., 2025; Kolluri et al., 2025).

3.1 Persona-conditioned Simulation

Let Q be a set of survey questions. Each question $q \in Q$ has answer options $X_q = \{x_1, \ldots, x_{N_q}\}$. For a demographic group g, defined by a single attribute–value pair (e.g., `AGE=30-49`), empirical human responses induce a subgroup distribution $U_{q,g}$ over X_q. Our goal is to produce a simulated distribution $S_{q,g}$ that is closely aligned with $U_{q,g}$.

To improve $S_{q,g}$, we condition each agent on a synthetic persona. Our framework proceeds in two stages: 1) persona generation and 2) persona-conditioned survey simulation. In the first stage, we induce a pool of personas from behavioral data. Each persona is a natural-language description of the individual's characteristics and is associated with inferred demographic attributes accompanied by corresponding confidence scores (see Figure 2), enabling it to be matched to a target demographic group g. In the second stage, these induced personas condition individual agents to simulate responses for a target survey, and the resulting predictions are aggregated for each demographic group.

For a target question q and group g, we select personas that match the demographic attribute defining g, prompt N_A LLM agents—one per persona—to produce a *verbalized distribution* (Hu et al., 2026)

over the answer options, and then aggregate outputs: [3] $S_{q,g}(x) = \frac{1}{N_A}\sum_{j=1}^{N_A} S_{q,g}^{(j)}(x), \quad x \in X_q$. This formulation separates persona construction from downstream evaluation, allowing the same simulation protocol to be instantiated with different persona sources and different levels of access to target-domain information.

3.2 Data Access Regimes

We organize our experiments based on the information available for persona construction and for assigning personas to demographic groups. Each data-access regime motivates a corresponding research question.

Low-access / out-of-domain. This regime represents situations where practitioners have no access to target-survey data. Personas must instead be induced from external sources, such as non-survey behavioral data, survey data from other datasets or waves, or the LLM's parametric knowledge. It mimics scenarios where only general behavioral evidence is available, testing whether such out-of-domain personas can still produce meaningful predictions for a target survey. This motivates *RQ1: Can personas induced from out-of-domain behavioral data improve survey simulation alignment?*

Mid-access / out-of-domain setting with GT demographics This regime reflects settings where practitioners have behavioral data from external sources along with GT demographic information. Personas can be induced using both behavioral traces and demographic attributes. This scenario captures cases where demographic alignment is feasible even without direct access to the target survey, allowing us to study how correct persona–group assignment affects simulation alignment. This motivates *RQ2: How does access to ground-truth demographic information in out-of-domain behavioral data improve survey simulation alignment?*

High-access / in-domain. This regime represents scenarios where practitioners have partial access to the target survey itself. Personas are induced from these in-domain responses and used to predict held-out questions within the same survey. It captures the potential of leveraging partial in-domain data to improve simulation for unseen questions, motivating *RQ3: Can in-domain personas induced from partial survey responses generalize to predict held-out questions?*

4 Experimental Setup

We now instantiate the study formulation described in Section 3 with concrete target datasets, persona sources, baselines, and evaluation metrics.

4.1 Target Survey Datasets Construction

We evaluate survey simulation on five publicly available survey datasets or waves: OpinionQA Waves 26 and 29 (Santurkar et al., 2023), the General Social Survey (GSS) (Davern et al., 2024), and two SocSci210 subsets, MP3DR and NJ5DX (Kolluri et al., 2025). [4] Each dataset provides respondent-level demographic attributes and individual survey responses, enabling construction of group-level empirical response distributions.

Following Section 3, we construct demographic group-level test instances by aggregating responses from respondents who share a given demographic attribute value. This yields test instances of the form $(q_i, \{x_k\}, g, U_{i,g})$. [5]

4.2 Persona Sources

We construct personas using three types of data.

Model knowledge. Given a target attribute defining the demographic group g, we prompt Claude 3.7 Sonnet (Anthropic, 2025) to generate a diverse set of personas consistent with that group. This provides a source-free reference point that relies only on the model's parametric knowledge.

Non-survey behavioral data. We induce personas from anonymized movie review histories (Yoon et al., 2024) and e-commerce shopping histories (Berke et al., 2024). This instantiates the low-access setting where only external out-of-domain behavioral traces are available.

Survey behavioral data. Personas are generated from anonymized respondent-level survey responses. When the source survey differs from the target, this represents an out-of-domain setting. When the source survey is the same as the target survey, we split the questions into persona-induction

[3] Based on the findings of our pilot study (Appendix A), we set $N_A = 30$, select the highest-confidence matching personas for each demographic group, and use verbalized distributions to represent agent outputs.

[4] OpinionQA waves and SocSci210 subsets differ in topics and populations and are therefore treated as separate surveys.

[5] Appendix D covers statistics of both target dataset and persona-source datasets and shows example of test instance.

and evaluation subsets, using the former as the behavioral data source and the latter as held-out test questions, representing the in-domain setting.

4.3 Persona Generation and Selection

For data-driven persona generation, we use Claude 3.7 Sonnet (Anthropic, 2025) as the persona generator. As shown in Figure 2, given an individual's behavioral record, the model produces: (i) a natural-language persona for simulation, and (ii) inferred demographic attributes for 11 keys (CREGION, AGE, SEX, EDUCATION, MARITAL, RELIG, RELIGATTEND, POLPARTY, POLIDEOLOGY, INCOME, and RACE) with confidence scores. Each key has a fixed closed set of options (e.g., AGE: 18–29 | 30–49 | 50–64 | 65+; POLPARTY: Democrat | Republican | Independent | Other); full option sets are listed in Appendix B.

For each demographic group g, we assign personas using their inferred demographic attributes, then rank candidates by confidence score and select the top $N_A = 30$ per group for simulation. This confidence-based selection is motivated by our pilot study (Appendix A), which shows that high-confidence extraction yields substantially higher demographic accuracy (e.g., 0.81 vs. 0.39 on GSS; Table 7), and that simulation performance largely saturates beyond 30 personas.

For settings where ground-truth demographic labels are available for source individuals *(Mid access / out-of-domain with GT demographics)*, we consider two additional variants. In *GT assignment*, ground-truth demographics are used only for the assignment of personas to different demographic groups. In *GT generation*, ground-truth demographic information is also provided as an additional data source during persona induction.

4.4 Baselines and Simulation Protocol

We compare the persona-conditioned simulation against three baselines following SimBench (Hu et al., 2026). [6]

B1 (No demographic context). Each agent receives only the question and answer options.

B2 (Country-level context). Each agent receives a short demographic context (e.g. *You are from USA*) in addition to the question.

[6]All baselines use the same individual-agent simulation and group-level aggregation procedure.

B3 (Single-attribute context). Each agent receives an explicit value of a demographic attribute (e.g. *You are Male*) along with the question.

Simulation protocol. For each target question q and demographic group g, we prompt $N_A = 30$ LLM agents in parallel, each conditioned on a distinct persona selected as described in Section 4.3. Each agent produces a verbalized probability distribution over the answer options (e.g., `A: 30%, B: 40%, C: 30%`), and the resulting distributions are aggregated via simple averaging as described in Section 3. This aggregation approximates within-group heterogeneity by pooling over a diverse set of persona-conditioned agents rather than collapsing each group to a single representative response.

4.5 Evaluation Metrics

We quantify the misalignment between the human demographic group-level response distribution $U_{i,g}(x)$ and the simulated response distribution $S_{i,g}(x)$ using two complementary metrics: Total Variation Distance (TVD) (Hu et al., 2026; Łajewska et al., 2026) and 1-Wasserstein distance (WD) (Suh et al., 2025; Moon et al., 2024) (see Appendix C for details). For each test instance (q_i, g), we compute the distance between $U_{i,g}$ and $S_{i,g}$ and report the average across all questions and demographic groups. Lower values indicate closer alignment for both metrics.

5 Results

We evaluate persona-conditioned simulation across five target survey datasets. [7] For each persona source dataset, we first construct a persona pool, then condition simulation on individual personas, and evaluate performance at the demographic group level. [8]

5.1 Low-access Regime / Out-of-domain Setting

We begin with a practically motivated low-access setting in which only human behavior traces from non-target sources are available (RQ1). We consider persona sources including external behavioral corpora (movie reviews and shopping histories) and survey respondent data from other datasets/waves.

[7]The simulation is executed with Claude 3.7 Sonnet using $N_A = 30$ agents for each demographic group. We provide additional results with open-weight LLMs in Appendix G.

[8]Note that the same persona pool can be reused for different target datasets in the out-of-domain setting.

Baseline/Persona Source	GSS		OpinionQA W26		OpinionQA W29		SocSci210 (MP3DR)		SocSci210 (NJ5DX)	
	TVD	WD	TVD	WD	TVD	WD	TVD	WD	TVD	WD
B1	0.169	0.322	0.189	0.299	0.171	0.262	<u>0.311</u>	**0.474**	0.241	0.678
B2	<u>0.160</u>	<u>0.301</u>	**0.173**	**0.274**	<u>0.139</u>	<u>0.204</u>	0.314	0.552	0.204	0.560
B3	**0.151**	**0.280**	<u>0.178</u>	<u>0.287</u>	**0.130**	**0.194**	0.323	0.583	0.204	0.538
Model Knowledge	<u>0.160</u>	0.303	0.190	0.311	0.157	0.237	0.330	0.579	**0.191**	**0.443**
Movie Reviews	0.169	0.326	0.197	0.318	0.176	0.263	0.315	0.592	<u>0.203</u>	<u>0.506</u>
Shopping History	0.163	0.312	0.189	0.301	0.160	0.240	**0.309**	<u>0.532</u>	0.208	0.530
SocSci210 MP3DR	0.186	0.352	0.203	0.324	0.168	0.248	–	–	0.251	0.746
SocSci210 NJ5DX	0.192	0.366	0.204	0.331	0.172	0.247	0.364	0.781	–	–
GSS	–	–	0.204	0.332	0.149	0.222	0.335	0.614	0.213	0.574
OpinionQA W26	0.167	0.320	–	–	0.151	0.229	0.320	0.608	0.205	0.533
OpinionQA W29	0.184	0.346	0.195	0.315	–	–	0.351	0.760	0.209	0.548

Table 1: Survey simulation results on the target survey datasets in the low data-access regime. The best-performing simulation approach for each target survey is indicated in **bold**, with the second-best <u>underlined</u>.

As shown in Table 1, simulation performance varies across target survey datasets, partly reflecting differences in the average number of response options (SocSci210 subsets have an average of 7 response options and correspondingly result in highest TVD, whereas GSS has an average of 3.1 options and yields the lowest TVD). In addition, variation in survey topics may further contribute to performance differences. [9]

Personas induced from diverse out-of-domain sources generally do not outperform simple baselines that provide minimal context about target demographic groups. In particular, B3, which conditions each agent on a single attribute at inference time, remains consistently competitive. Notably, personas derived from sources such as shopping histories and movie reviews are also not particularly effective. One possible explanation is that persona sources often focus on relatively narrow topics (e.g., OpinionQA W29 centers on gender-related issues, while movie reviews reflect domain-specific preferences), leading to personas that are overly specific and misaligned with the target surveys. This suggests that, without strong alignment between source behavioral data and the target population, additional persona detail may introduce noise rather than improve simulation alignment. Overall, the results indicate that group-level response distributions can often be approximated from minimal demographic context, and that naively incorporating out-of-domain personas does not reliably add predictive signal.

These results should also be interpreted in light of what distribution alignment does not capture. Even when persona-conditioned simulation does not substantially improve simulation alignment, it can still offer practical advantages in interpretability and explainability. Because each simulated response is tied to a particular persona induced from behavioral data, the resulting predictions are easier to inspect and explain than those from single-attribute baseline. Moreover, using a pool of personas enables the simulator to capture variation within a demographic group, rather than representing that group with a single-attribute demographic information.

We attribute the limited gains from out-of-domain personas to two primary factors. First, behavioral data from a source dataset may contain signals that are not useful for simulating users' opinions in target dataset. Second, our pipeline maps personas to target groups using demographic attributes inferred from behavioral data; inaccurate demographic attribute inference can therefore cause persona–group misassignment (e.g., assigning a persona derived from behavioral data of a female user to a male target demographic group because of incorrect demographic feature extraction), degrading simulation alignment. [10] These observations motivate RQ2.

5.2 Mid-access Regime / Out-of-domain Setting with GT Demographics

For datasets where individual-level GT demographic attributes are available, we evaluate two

[9] TVD and WD generally produce similar method rankings; unless otherwise noted, we therefore base our analysis on TVD.

[10] Details of demographic feature inference accuracy are provided in Tables 7 and 8 in Appendix F.

Baseline/Persona Source	GSS		OpinionQA W26		OpinionQA W29		SocSci210 (MP3DR)		SocSci210 (NJ5DX)	
	TVD	WD	TVD	WD	TVD	WD	TVD	WD	TVD	WD
B3	**0.151**	**0.280**	0.178	0.287	**0.130**	**0.194**	0.323	0.583	0.204	0.538
GSS	–	–	0.204	0.332	0.149	0.222	0.335	0.614	0.213	0.574
+GT assignment	–	–	0.200	0.325	0.141	0.210	0.319	0.512	0.194	0.474
+GT generation	–	–	0.208	0.336	0.149	0.236	0.317	**0.482**	0.200	0.495
OpinionQA W26	0.167	0.320	–	–	0.151	0.229	0.320	0.608	0.205	0.533
+GT assignment	0.154	0.304	–	–	0.141	0.214	**0.305**	0.544	0.192	0.472
+GT generation	0.157	0.291	–	–	0.150	0.234	0.329	0.571	**0.186**	0.457
OpinionQA W29	0.184	0.346	0.195	0.315	–	–	0.351	0.760	0.209	0.548
+GT assignment	0.159	0.304	0.172	0.274	–	–	0.326	0.648	0.188	**0.447**
+GT generation	0.159	0.293	**0.169**	**0.267**	–	–	0.348	0.677	0.187	0.450

Table 2: Survey simulation results on the target survey datasets in mid data-access regime. The best-performing approach for each survey is indicated in **bold**, with the second-best underlined. *GT assignment* uses GT demographics only for persona–group assignment, while *GT generation* uses GT demographics during persona generation.

strategies for leveraging GT to improve persona-conditioned survey simulation: (i) *GT assignment*, where GT demographic information is used only to improve the assignment of personas to target demographic groups, and (ii) *GT generation*, where persona generation is conditioned on both behavioral data and the corresponding GT demographic information about the user.

As shown in Table 2, access to GT demographics substantially improves persona-conditioned simulation. GT assignment consistently reduces TVD relative to using the same persona source without GT demographic information, indicating that demographic attribute misassignment is a major contributor to reduced cross-domain simulation alignment. In contrast, GT generation yields mixed effects: it can further reduce TVD in some settings, but it is not uniformly better than GT assignment, suggesting that conditioning persona generation on demographic information can introduce additional variance depending on the source behavioral data.

Our qualitative analysis further suggests that conditioning persona generation on GT demographics can make the resulting personas disproportionately focused on surface-level attributes, rather than higher-level traits such as interests, preferences, or behavioral tendencies (see Section H.1). Overall, these findings highlight accurate demographic assignment as a key bottleneck: when personas are correctly assigned to target demographic groups, persona-conditioned simulation becomes substantially more aligned to human responses even when the source behavioral data differs from the target survey population.

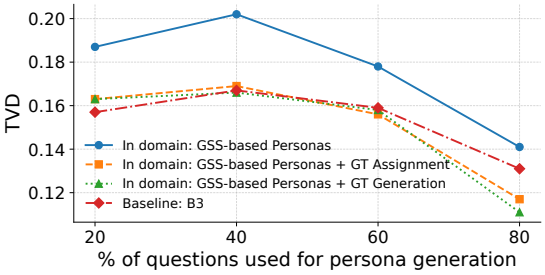

Figure 3: Survey simulation results in the in-domain setting, showing TVD for simulations using personas generated from varying fractions of GSS questions.

5.3 High-access Regime / In-domain Setting

We investigate how well personas induced from target-survey data generalize within the same domain (RQ3). We use the GSS dataset for this setting, as it covers diverse topics and includes enough questions to allow meaningful splits. Questions are divided into two subsets, with {20%, 40%, 60%, 80%} used for persona generation and the remainder for simulation evaluation.

As shown in Figure 3, simulation performance improves as more question history is used, with the largest gains at 80%. One interpretation is that additional questions provide richer behavioral evidence, enabling the persona generator to infer more stable attributes, such as preferences or attitudes toward topics related to the held-out questions, which in turn improves response simulation. Moreover, when personas incorporate a larger portion of each respondent's history, the simulator can better capture consistent response patterns across related questions. This leads to more faithful simulations of the respondent represented by each per-

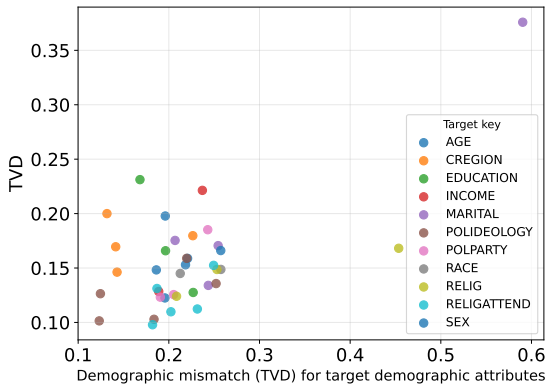

Figure 4: For each target demographic group, we measure the correlation between demographic mismatch score (as measured by TVD between selected personas and actual demographics) and survey TVD. Correlation coefficient: (Pearson=0.631, Spearman = 0.323)

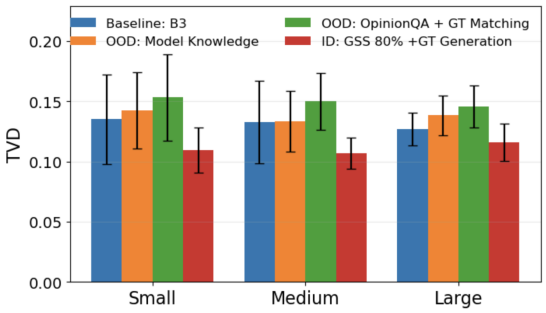

Figure 5: TVD across different demographic-group sizes after binning groups into Small, Medium, and Large buckets by respondent count in GSS dataset.

sona and, consequently, better alignment between human and agent response distributions within specific demographic groups.

In-domain personas with GT generation consistently outperforms the baseline once 40% or more of the question history is used. This result suggests that personas generalize effectively across different question sets within the same domain, and further indicates that access to a sufficient amount of behavioral data from the target source is important for improving survey simulation.

To verify that this finding is not an artifact of a single data split, we repeat the experiment across multiple splits. GT Generation outperforms B3 in all three independent runs at the 80% split (0.111 vs. 0.131; 0.139 vs. 0.147; 0.153 vs. 0.169). While marginal standard deviations overlap between conditions (see Figure 7 for full results including standard deviations), this paired comparison confirms that the observed ranking is stable across splits. Due to computational cost, remaining conditions in Tables 1, 2, and 5 are run once; we treat this repeated-run result as verification that our pipeline produces reproducible relative orderings.

6 Discussion

In this section, we provide deeper analyses of our persona-conditioned simulation approaches and baselines. Unless stated otherwise, all analyses adopt the GSS evaluation configuration in which personas are induced from 80% of the survey questions and evaluated on the remaining 20% (as described in Section 5.3). We focus on representative methods from each investigated data-access regime spanning the single-attribute context baseline (B3), model-knowledge personas (Model Knowledge), out-of-domain personas with GT assignment sourced from OpinionQA Wave 26 (OpinionQA), and in-domain personas induced from 80% of the question history with GT demographic information (GSS 80%).

6.1 Effect of Persona Representativeness on Simulation Alignment

We investigate whether simulation performance depends on how well the selected personas capture the ground-truth respondent population beyond the target attribute. Under the GSS 80% setting, we compare, for each target subgroup (e.g., `Gender=male` or `Age=18-29`), the 30 personas used for simulation with the corresponding respondent subgroup. For each non-target demographic attribute, we compute TVD between the attribute distribution across the selected personas and the distribution in the corresponding respondent group, considering all possible attribute values (e.g., education levels or income brackets). This yields a separate TVD score for each non-target attribute, which we then average to obtain an overall *demographic mismatch* score for the demographic group.

As shown in Figure 4, we observe a moderate positive association between demographic mismatch and simulation misalignment (TVD) (Pearson $= 0.631$, Spearman $= 0.323$), suggesting that demographic groups for which selected personas more closely resemble the corresponding respondent population tend to exhibit lower simulation misalignment. The weaker Spearman correlation, however, indicates that the relationship is not uniformly monotonic across demographic groups.

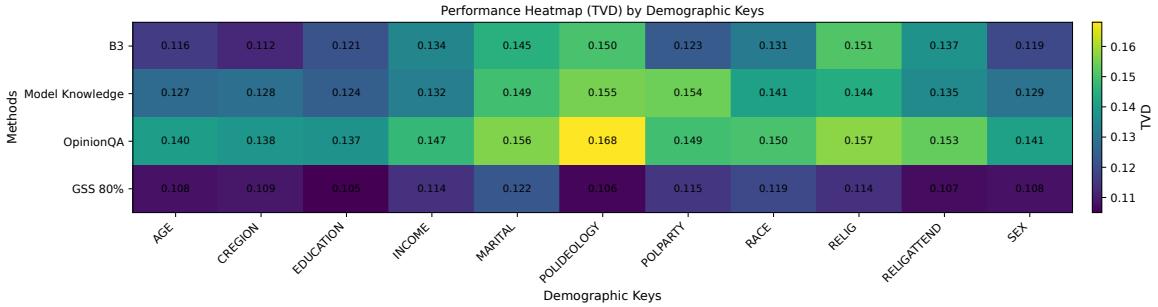

Figure 6: Performance Heatmap (TVD) across demographic keys

6.2 Simulation Alignment Variations across Demographic Groups

Group size. We first compare performance across demographic-group sizes to assess whether simulation remains effective for relatively underrepresented groups across different data-access regimes.[11]

As shown in Figure 5, the ranking of methods is stable across all group sizes: personas induced from 80% of GSS questions consistently achieve the best simulation alignment, followed by B3 and Model Knowledge, while OpinionQA exhibits the highest TVD. This suggests that the benefits of in-domain persona induction are not limited to well-populated groups but extend to both sparse and large groups. The advantage is especially pronounced for smaller groups, where underrepresentation presents a greater challenge. Overall, these results suggest that in-domain persona induction yields stronger and more stable alignment across groups of varying sizes.

Demographic key. We next compare performance across demographic keys (e.g., `AGE` and `REGION`) to examine which attributes are more consistently captured by each method. As shown in Figure 6, GSS 80% performs best across all demographic attributes, with uniformly low TVD values. This indicates that personas induced from in-domain evidence generalize reliably across attributes, rather than improving only a narrow subset of groups. Among the remaining methods, B3 is generally the strongest baseline, particularly on age, education, race, and sex, where differences are relatively small. By contrast, OpinionQA consistently underperforms, indicating limited cross-domain simulation alignment when the source distribution diverges from the target. Model Knowledge performs between B3 and OpinionQA, and overall shows notable degradation on POLPARTY (0.154), suggesting that political affiliation is especially difficult to simulate from model knowledge alone when the target distribution differs from the model's internal knowledge.

[11]The number of respondents in GSS does not fully reflect the true size of each demographic group, but it can serve as a proxy for population representation (e.g., smaller groups may be viewed as relatively underrepresented groups).

7 Conclusion

We presented a systematic study of persona-conditioned survey simulation under different data-access regimes, evaluated at the demographic-group level. Our results show that out-of-domain personas rarely outperform a single-attribute context baseline, primarily due to domain mismatch and errors in demographic attribute assignment. Their effectiveness improves substantially when ground-truth demographics are available, particularly for persona–group assignment. In-domain personas perform progressively better as more question history is incorporated, suggesting that richer behavioral evidence enables more stable trait inference and better generalization to held-out questions. Overall, these findings highlight a practical insight for survey simulation: the critical factor is not merely the use of personas, but ensuring they are grounded in sufficient behavioral evidence and accurately matched to the target population.

Limitations

First, our evaluation has several scope limitations: it focuses on single-attribute demographic groups rather than intersectional groups, which may understate the difficulty of simulating finer-grained subpopulations; it is limited to U.S. respondents because our behavioral data sources are U.S.-centric; and it evaluates distribution-level alignment rather than individual-level, prioritizing privacy-aligned

deployment. Extending the evaluation to intersectional groups and other regions, and connecting distributional accuracy to downstream decision-making and application-specific utility, remain important directions for future work.

Second, we did not investigate methods for improving demographic feature extraction accuracy. Table 8 shows that extraction errors are a major bottleneck in the low-access regime, as they lead to persona misassignment and degrade simulation alignment. Our mid-access results suggest that accurate demographic assignment is the key factor, providing a concrete target for future work on more robust demographics inference methods.

Finally, due to computational cost, most experiments are run only once. Rather than focusing on stability through repeated runs, we prioritize generalizability by evaluating across 5 target datasets, 8 persona sources, and 3 LLMs, where consistent trends are observed. Since this is the first work to characterize persona generation from diverse behavioral data domains and systematically measure persona fidelity across domains, we prioritize generalization over stability. For the in-domain setting, where uncertainty estimation is most relevant given the varying question history splits, we conduct multiple runs and report standard deviation in Figure 7 and Table 6.

Ethical Considerations

In our experiments, we utilize publicly available datasets, including OpinionQA (Santurkar et al., 2023), GSS (Davern et al., 2024), SocSci210 (Kolluri et al., 2025), movie review (Yoon et al., 2024), and shopping histories (Berke et al., 2024). These datasets are widely adopted and well-established within the research community, and their use raises no additional privacy or consent concerns beyond those addressed in the original dataset releases.

All large language models used in this work were accessed through their official and publicly available sources. Our use of these models complies with their stated terms of service and aligns with open science and reproducibility principles.

During the preparation of this manuscript, we employed an AI-assisted writing tool at the sentence level to support drafting and linguistic refinement.

Beyond the privacy safeguards already described, we note additional societal risks specific to persona-conditioned survey simulation. First, tools that simulate group-level opinion distributions could be misused for political microtargeting or opinion manipulation. Second, survey simulation risks being used as a substitute for genuine stakeholder engagement, particularly for underrepresented or hard-to-reach populations. Our results suggest this is especially risky in the low-access regime, where simulated distributions can appear plausible while diverging substantially from real subgroup opinions due to demographic misassignment (Section 5.1). We emphasize that our findings argue against using out-of-domain persona simulation as a drop-in replacement for genuine data collection, and recommend that any deployment report demographic-assignment accuracy and group-level sample size alongside simulated distributions.

References

Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *International conference on machine learning*, pages 337–371. PMLR.

Anthropic. 2025. Claude 3.7 sonnet and claude code.

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.

Christopher A Bail. 2024. Can generative ai improve social science? *Proceedings of the National Academy of Sciences*, 121(21):e2314021121.

Ziyad Benomar, Weronika Łajewska, Leonardo Perelli, and Saab Mansour. 2026. Data-driven persona-conditioned agents for A/B test simulation. *arXiv preprint arXiv:2609.01038*.

Alex Berke, Dan Calacci, Robert Mahari, Takahiro Yabe, Kent Larson, and Sandy Pentland. 2024. Open e-commerce 1.0, five years of crowdsourced us amazon purchase histories with user demographics. *Scientific Data*, 11(1):491.

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. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3141–3154.

Serina Chang, Alicja Chaszczewicz, Emma Wang, Maya Josifovska, Emma Pierson, and Jure Leskovec. 2025. Llms generate structurally realistic social networks but overestimate political homophily. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pages 341–371.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, and 1 others. 2024. From persona to personalization: A survey on role-playing language agents. *arXiv preprint arXiv:2404.18231*.

Myra Cheng, Tiziano Piccardi, and Diyi Yang. 2023. Compost: Characterizing and evaluating caricature in llm simulations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10853–10875.

Michael Davern, Rene Bautista, Jeremy Freese, Pamela Herd, and Stephen L. Morgan. 2024. General social survey 2024. [Machine-readable data file]. Principal Investigator: Michael Davern; Co-Principal Investigators: Rene Bautista, Jeremy Freese, Pamela Herd, and Stephen L. Morgan. Sponsored by National Science Foundation. Data accessed from the GSS Data Explorer.

Ricardo Dominguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dünner. 2024. Questioning the survey responses of large language models. *Advances in Neural Information Processing Systems*, 37:45850–45878.

Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. 2024. Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanities and Social Sciences Communications*, 11(1):1259.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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

John J Horton. 2023. Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research.

Tiancheng Hu, Joachim Baumann, Lorenzo Lupo, Nigel Collier, Dirk Hovy, and Paul Röttger. 2026. Simbench: Benchmarking the ability of large language models to simulate human behaviors. In *International Conference on Learning Representations*, volume 2026, pages 28290–28341.

EunJeong Hwang, Bodhisattwa Majumder, and Niket Tandon. 2023. Aligning language models to user opinions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5906–5919.

Shivani Kapania, William Agnew, Motahhare Eslami, Hoda Heidari, and Sarah E Fox. 2025. Simulacrum of stories: Examining large language models as qualitative research participants. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–17.

Jaehyung Kim and Yiming Yang. 2025. Few-shot personalization of llms with mis-aligned responses. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11943–11974.

Junsol Kim and Byungkyu Lee. 2023. Ai-augmented surveys: Leveraging large language models and surveys for opinion prediction. *arXiv preprint arXiv:2305.09620*.

Akaash Kolluri, Shengguang Wu, Joon Sung Park, and Michael S Bernstein. 2025. Finetuning llms for human behavior prediction in social science experiments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30084–30099.

Louis Kwok, Michal Bravansky, and Lewis D Griffin. 2024. Evaluating cultural adaptability of a large language model via simulation of synthetic personas. *arXiv preprint arXiv:2408.06929*.

Weronika Łajewska, Paul Missault, George Davidson, and Saab Mansour. 2026. RADIUS: Ranking, distribution, and significance-a comprehensive alignment suite for survey simulation. *arXiv preprint arXiv:2603.19002*.

Benjamin S Manning, Kehang Zhu, and John J Horton. 2024. Automated social science: Language models as scientist and subjects. Technical report, National Bureau of Economic Research.

Saab Mansour, Leonardo Perelli, Lorenzo Mainetti, George Davidson, and Stefano D'Amato. 2025. Paars: Persona aligned agentic retail shoppers. In *Proceedings of the 1st Workshop for Research on Agent Language Models (REALM 2025)*, pages 143–159.

Nicole Meister, Carlos Guestrin, and Tatsunori B Hashimoto. 2025. Benchmarking distributional alignment of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 24–49.

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. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 19864–19897.

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

Peter S Park, Philipp Schoenegger, and Chongyang Zhu. 2024b. Diminished diversity-of-thought in a standard large language model. *Behavior Research Methods*, 56(6):5754–5770.

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. 2023. Moral mimicry: Large language models produce moral rationalizations tailored to political identity. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 282–297.

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. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21147–21170.

Chenkai Sun, Ke Yang, Revanth Gangi Reddy, Yi R Fung, Hou Pong Chan, Kevin Small, Cheng Xiang Zhai, and Heng Ji. 2025. Persona-db: Efficient large language model personalization for response prediction with collaborative data refinement. In *31st International Conference on Computational Linguistics, COLING 2025*, pages 281–296. Association for Computational Linguistics (ACL).

Seungjong Sun, Eungu Lee, Dongyan Nan, Xiangying Zhao, Wonbyung Lee, Bernard J Jansen, and Jang Hyun Kim. 2024. Random silicon sampling: Simulating human sub-population opinion using a large language model based on group-level demographic information. *arXiv preprint arXiv:2402.18144*.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two tales of persona in llms: A survey of role-playing and personalization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16612–16631.

Lei Wang, Jingsen Zhang, Hao Yang, Zhi-Yuan Chen, Jiakai Tang, Zeyu Zhang, Xu Chen, Yankai Lin, Hao Sun, Ruihua Song, and 1 others. 2025. User behavior simulation with large language model-based agents. *ACM Transactions on Information Systems*, 43(2):1–37.

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*.

Se-eun Yoon, Zhankui He, Jessica Echterhoff, and Julian McAuley. 2024. Evaluating large language models as generative user simulators for conversational recommendation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1490–1504.

Persona Source				
Dataset	**Domain**	**#Q**	**Avg. Resp.**	**#Resp.**
GSS	General Survey	60	36.1	3309
OpinionQA W26	Survey: Guns	78	46.2	4168
OpinionQA W29	Survey: Gender	77	53.8	4867
SocSci210 MP3DR	Survey: Gender	39	6.9	4161
SocSci210 NJ5DX	Survey: Inequality	48	18.2	1814
IMDB	Movie review	–	21.2	1079
E-commerce	Shopping history	–	368.2	5027
Test Dataset				
Dataset	**#Demog. feats.**	**#Q**	**#Opt.**	**Total**
GSS	40	60	3.1	2280
OpinionQA W26	50	78	4.5	3474
OpinionQA W29	50	77	4.1	3622
SocSci210 MP3DR	18	39	7.0	555
SocSci210 NJ5DX	25	44	7.0	934

Table 3: Summary statistics of persona sources and test datasets.

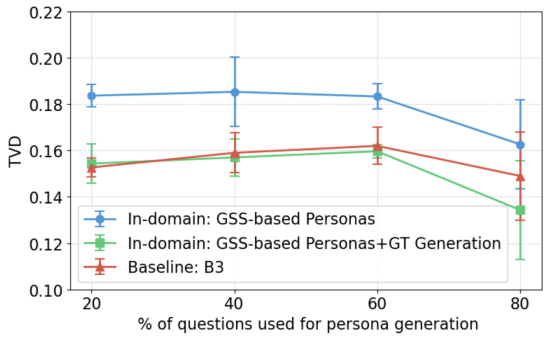

Figure 7: Survey simulation results in the in-domain setting, showing TVD and standard deviation (Std.) for simulations using personas generated from varying fractions of GSS questions. Results for each iteration and Std. are available in Table 6

A Pilot Study

In our pilot study, we explore several design choices for using synthetic personas in survey-response simulation.

A.1 Verbalized Distribution vs. Option Selection

Section 3.1 elicits a *verbalized* probability distribution from each persona-conditioned agent (e.g., `A: 30%, B: 70%`) and then averages these distributions. As an alternative, we can estimate the group-level distribution by prompting each agent to select *a single* answer option from the available set and then aggregating these selections across agents.

Concretely, for each agent $j \in \{1, \ldots, N_A\}$ conditioned on persona P_j, we prompt the LLM to output exactly one option $\hat{x}_q^{(j)} \in \mathcal{X}_q$ for question q. We then form an empirical distribution by counting how often each option is selected:

$$S_{q,g}(x) = \frac{1}{N_A} \sum_{j=1}^{N_A} \mathbb{1}\left\{\hat{x}_q^{(j)} = x\right\}, \quad x \in \mathcal{X}_q, \tag{1}$$

where $\mathbb{1}\{\cdot\}$ denotes the indicator function.

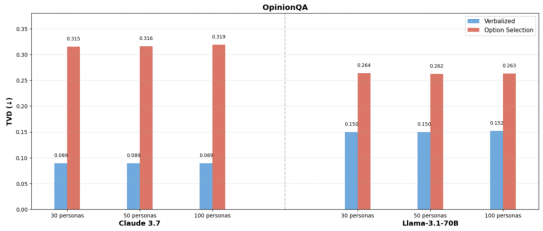

Figure 8: Performance of Claude-3.7-Sonnet and Llama-3.1-70B model on OpinionQA across different number of personas and different distribution aggregation methods

A.2 Number of Personas and Aggregation Scheme: Verbalized Distributions vs. Option Selection

We compare two aggregation schemes: (i) *verbalized distribution* outputs, in which the simulator directly produces a probability distribution over answer options, and (ii) *option selection + aggregation*, in which each persona selects a single option and responses are aggregated by voting. We also vary the number of personas used for each demographic group.

We conduct experiments on SimBench-OpinionQA (Hu et al., 2026) using Claude 3.7 Sonnet and Llama-3.1-70B as simulators. For each target demographic value, we evaluate simulations using 30, 50, and 100 personas, and compare the two aggregation schemes.

As shown in Figure 8, verbalized distributions consistently outperform option selection + aggregation. This finding aligns with prior work suggesting that direct distribution elicitation better preserves minority probability mass than hard voting (Meister et al., 2025). We also find that performance largely saturates beyond 30 personas: once the number of personas exceeds approximately 30, the improvement curve plateaus, suggesting diminishing returns from adding more agents under this setup.

A.3 Why Does Option Selection Underperform?

We next investigate why the option selection scheme performs worse than verbalized distri-

GT Generation	Free
You are a 33-year-old Black man living in the Northeast who works in education administration at an elementary or secondary school. You're a college graduate with a bachelor's degree who has never married but has two children. You identify as Protestant and attend religious services weekly, indicating that faith is an important part of your life. Politically, you lean Democratic but consider yourself moderate rather than strongly liberal. You consistently vote in presidential elections, supporting Democratic candidates like Clinton and Biden. You have only some confidence in the executive branch of government, suggesting a degree of political skepticism despite your party affiliation...	You are a politically moderate Democrat who consistently votes in elections and supports Democratic presidential candidates. While you identify as middle-of-the-road politically, your policy positions tend to lean progressive on domestic issues - you support increased spending on education, healthcare, environmental protection, addressing urban problems, and improving conditions for Black Americans. You oppose the death penalty and believe premarital sex is not morally wrong. You're fiscally more conservative on international matters, believing we spend too much on military and foreign aid. You're civically engaged and vote regularly in presidential elections...

Table 4: GT feature vs. Free description

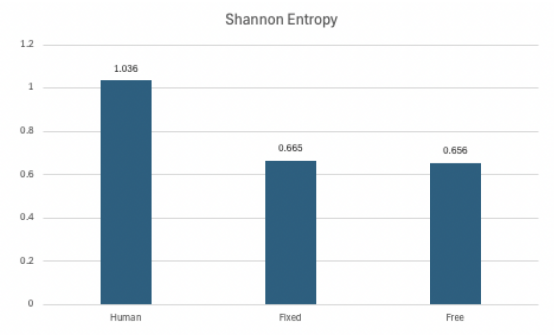

Figure 9: Average Shannon entropy of simulated response distributions under the optionselection aggregation scheme. Lower entropy indicates more peaked (less diverse)distributions, showing that option selection tends to collapse onto a single dominant option.

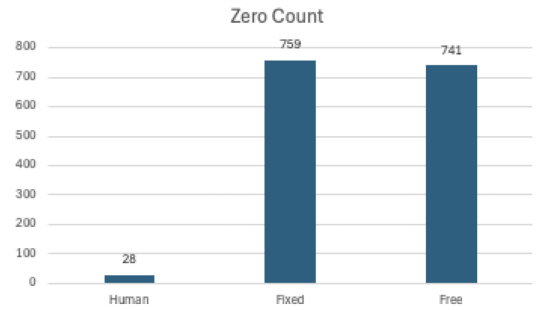

Figure 10: Number of answer options with zero counts under the option selection aggregation scheme, aggregated across survey questions. A high number of zero-count options indicates that minority opinions are rarely represented in the simulated distributions.

butions. Our working hypothesis is that majority voting over single-choice responses produces overly peaked distributions that over-represent the dominant preference while systematically under-representing minority opinions.

To examine this hypothesis, we compute two diagnostics: (1) the average Shannon entropy of the simulated response distribution, where lower entropy indicates a more concentrated distribution, and (2) the number of non-chosen options across questions, where for each of the 964 survey questions we count how many answer options receive zero selections under option selection + aggregation.

Across questions, we find that option selection frequently collapses to a single dominant option, with most of the probability mass concentrated on one answer and many alternatives never selected even once. As a result, the simulated distribution primarily reflects the majority view and fails to capture the within-group heterogeneity observed in human data.

This effect is further amplified by the persona construction used in this pilot setting: personas are generated from only a single demographic attribute (e.g., *Gender: female*). With such limited information to differentiate personas, agents tend to converge on the same "typical" response, and voting-based aggregation then amplifies this consensus while suppressing minority mass. This mechanism plausibly explains why option-selection-based distributions align less well with the true subgroup-level response distributions.

A.4 Persona Induction Strategies

To study how prompting affects persona quality, we compare four induction strategies:

1. **Free.** Generate a free-form bio from the individual's behavioral data only.

2. **Free+GT.** Generate a free-form bio from the individual's behavioral data + Ground-truth demographic feature as described in Section 5.2.

3. **Target-aware.** Provide brief context about the target survey datasets to be simulated (task description, domains, and example questions), then generate a bio.

4. **Rubric-based.** Ask the generator to answer survey-style rubric questions (aligned with target survey domains) based on the individual's data, then synthesize a bio.

Each persona includes (i) a short bio and (ii) extracted demographic features with confidence scores. Personas are then grouped by demographic features (e.g., Gender: male) and selected accordingly for subgroup simulation. In pilot experiments, we found that using the *Free* prompt template for persona generation and selecting personas with high confidence on the target demographic attribute consistently yielded the best performance. Detailed prompt templates used in the experiments (Free, Free+GT) are described in Figure 11 and 12, respectively.

B Persona Detail

Each key has a closed set of possible values:

- **Region**: closed options (South | Northeast | West | Midwest)
- **Age**: closed options (18–29 | 30–49 | 50–64 | 65+)
- **Sex**: closed options (male | female)
- **Education**: closed options (high school graduate | college graduate/some postgrad | some college, no degree | associate's degree | postgraduate | less than high school)
- **Marital**: closed options (married | divorced | widowed | living with a partner | never been married)
- **Religion**: closed options (protestant | roman catholic | atheist | agnostic | nothing in particular)
- **Religious Attendance**: closed options (more than once a week | seldom | never | a few times a year | once or twice a month | once a week)
- **Political Party**: closed options (Democrat | Independent | Other | Republican)
- **Income**: closed options (Less than $30,000 | $30,000–$50,000 | $50,000–$75,000 | $75,000–$100,000 | $100,000 or more)
- **Political Ideology**: closed options (liberal | moderate | conservative | very conservative | very liberal)
- **Race**: closed options (black | hispanic | asian | white)

C Metric Details

TVD measures the ℓ_1 discrepancy between two discrete distributions:

$$\mathrm{TVD}(U_{i,g}, S_{i,g}) = \frac{1}{2} \sum_{k=1}^{N_{\mathrm{option},i}} |U_{i,g}(x_k) - S_{i,g}(x_k)| \,. \tag{2}$$

WD captures the minimal "mass transport" needed to transform one distribution into another while respecting the ordering of options.[12]

$$W_1(U_{i,g}, S_{i,g}) = \sum_{k=1}^{N_{\mathrm{option},i}} \left| \sum_{r=1}^{k} U_{i,g}(x_r) - \sum_{r=1}^{k} S_{i,g}(x_r) \right| . \tag{3}$$

D Dataset Details

D.1 Statistics of Target Datasets and Persona-source Datasets

As discussed in Section 4, we utilize 5 different survey datasets for test and use various sources of persona induction. Table 3 summarize the statistics for datasets used in our experiments.

D.2 Group-level Test Instances Example

Figure 13 shows example test instances from each target survey datasets.

E Inference Details

For persona generation we used temperature = 0.7 and for simulation inference (both baseline and persona conditioned inference) we used temperature = 0.0

Prompt used for synthetic persona-conditioned simulation is described in Figure 14.

[12] This metric complements TVD by accounting for the ordinal structure of response options. All target datasets used in our experiments consist of survey questions with ordinal options, which are suited for WD.

F Accuracy of Demographic Feature Extraction in Persona Generation

In this section, we evaluate whether demographic features extracted from behavior data match ground-truth demographics. Ground-truth demographics are available for OpinionQA W26/W29 and GSS.

F.1 Extraction Accuracy by Confidence Level

We first analyze how extraction accuracy varies with the model-assigned confidence level. As shown in Table 7, accuracy usually increases with confidence across datasets, consistent with Appendix A's finding that high-confidence persona selection improves simulation performance.

We then test whether the personas used for simulation ($N_A = 30$ per demographic feature) are actually matched to the intended target group. The results shown in Table 8 indicate substantial mismatch, which helps explain the weaker performance of personas constructed without ground-truth demographic conditioning: when persona demographics are misaligned with the target subgroup, persona conditioning can become noisy or misleading.

G Additional Results

To test robustness of survey simulation across model families, we report results for two additional models (Llama-3.3-70B (Grattafiori et al., 2024) and Qwen-3-32B (Yang et al., 2025)) using personas induced from diverse sources in Table 5. The findings in Section 5.1 regarding persona utility hold consistently across models.

H Additional Discussions

H.1 Personas Induced from Behavioral Data Alone vs. with GT Demographics

Providing GT demographics during persona generation sometimes reduces TVD in certain settings, but it does not uniformly outperform GT assignment, as discussed in Section 5.2. Our manual inspection of these personas suggests a systematic difference in how they are constructed. Personas generated with GT demographics typically begin with an explicit listing of the provided demographic attributes, followed by a behavioral summary grounded in the respondent's observed answers. In contrast, personas generated without GT demographic information often express demographic characteristics only implicitly, or leave them underspecified, while relying more heavily on generic behavioral patterns (see Appendix Table 4 for an example).

These observations suggest that providing GT demographic information may encourage the generation of shallower personas that focus less on actual behavioral patterns. This may help explain why GT-informed persona generation does not consistently lead to improved simulation performance.

H.2 Broader Implications for Survey Simulation Systems

Taken together, our results suggest that the main value of synthetic personas for survey simulation is not simply that they provide richer prompt text, but that they can encode behaviorally grounded evidence when they are representative of the population being simulated. In practice, this means future survey simulation systems should treat persona use as an alignment problem rather than a prompting problem: they should validate how well personas represent source-population and explicitly audit persona–group matching. Our findings also suggest that when GT demographics are available, they are especially useful for matching, whereas injecting them directly into persona generation may encourage shallower, surface-level profiles. Finally, when partial target-survey history is available, systems should prioritize it, since richer in-domain behavioral evidence yields more stable traits, better transfer to unseen questions, and more consistent gains across demographic groups. The practical impact of this work is therefore to provide design guidance for building more reliable and privacy-aligned survey simulators using synthetic personas derived from diverse behavioral data.

Simulator: Claude 3.7										
Target Data	GSS		OpinionQA W26		OpinionQA W29		SocSci210 (MP3DR)		SocSci210 (NJ5DX)	
	TVD	WD	TVD	WD	TVD	WD	TVD	WD	TVD	WD
B1	0.169	0.322	0.189	0.299	0.171	0.262	0.311	0.474	0.241	0.678
B2	0.160	0.301	**0.173**	0.274	0.139	0.204	0.314	0.552	0.204	0.560
B3	**0.151**	0.280	0.178	0.287	**0.130**	0.194	0.323	0.583	0.204	0.538
Persona Source										
Model Knowledge	0.160	0.303	0.190	0.311	0.157	0.237	0.330	0.579	**0.191**	0.443
Movie	0.169	0.326	0.197	0.318	0.176	0.263	0.315	0.592	0.203	0.506
Shopping His	0.163	0.312	0.189	0.301	0.160	0.240	**0.309**	0.532	0.208	0.530
Survey MP3DR	0.186	0.352	0.203	0.324	0.168	0.248	–	–	0.251	0.746
Survey NJ5DX	0.192	0.366	0.204	0.331	0.172	0.247	0.364	0.781	–	–
OpinionQA W26	0.167	0.320	–	–	0.151	0.229	0.320	0.608	0.205	0.533
OpinionQA W29	0.184	0.346	0.195	0.315	–	–	0.351	0.760	0.209	0.548
GSS	–	–	0.204	0.332	0.149	0.222	0.335	0.614	0.213	0.574
Simulator: Llama-3.3-70B										
Target Data	GSS		OpinionQA W26		OpinionQA W29		SocSci210 (MP3DR)		SocSci210 (NJ5DX)	
	TVD	WD	TVD	WD	TVD	WD	TVD	WD	TVD	WD
B1	0.200	0.357	0.200	0.318	0.200	0.295	0.475	0.863	0.305	0.879
B2	**0.168**	0.320	**0.183**	0.287	0.188	0.279	0.456	0.861	0.325	1.030
B3	0.171	0.311	**0.183**	0.291	**0.181**	0.264	0.431	0.808	0.311	0.947
Persona Source										
Model Knowledge	0.169	0.320	0.185	0.297	0.195	0.294	0.427	0.687	**0.275**	0.803
Movie	0.191	0.362	0.210	0.341	0.239	0.366	0.472	0.868	0.315	0.983
Shopping His	0.180	0.342	0.212	0.345	0.208	0.317	0.441	0.820	0.277	0.864
Survey MP3DR	0.209	0.392	0.199	0.322	0.235	0.348	–	–	0.390	1.290
Survey NJ5DX	0.209	0.398	0.200	0.323	0.225	0.334	0.398	0.787	–	–
OpinionQA W26	0.192	0.355	–	–	0.208	0.309	0.453	0.832	0.283	0.860
OpinionQA W29	0.194	0.366	0.204	0.327	–	–	**0.384**	0.660	**0.275**	0.835
GSS	–	–	0.209	0.336	0.198	0.306	0.431	0.782	0.303	0.943
Simulator: Qwen3-32B										
Target Data	GSS		OpinionQA W26		OpinionQA W29		SocSci210 (MP3DR)		SocSci210 (NJ5DX)	
	TVD	WD	TVD	WD	TVD	WD	TVD	WD	TVD	WD
B1	0.186	0.337	0.228	0.381	0.227	0.350	0.595	1.674	0.247	0.710
B2	0.180	0.328	0.235	0.390	**0.189**	0.277	0.556	1.473	0.260	0.721
B3	0.178	0.319	0.221	0.369	0.194	0.282	0.547	1.486	0.254	0.716
Persona Source										
Model Knowledge	**0.173**	0.311	**0.207**	0.346	0.192	0.284	0.559	1.451	0.219	0.561
Movie	0.183	0.324	0.229	0.380	0.214	0.321	0.542	1.245	0.253	0.699
Shopping His	0.174	0.314	0.218	0.358	0.193	0.282	0.523	1.306	0.226	0.626
Survey MP3DR	0.199	0.366	0.228	0.376	0.212	0.310	–	–	0.284	0.846
Survey NJ5DX	0.203	0.372	0.222	0.373	0.217	0.315	0.592	1.594	–	–
OpinionQA W26	0.176	0.323	–	–	0.192	0.283	**0.492**	1.200	0.229	0.621
OpinionQA W29	0.184	0.341	0.217	0.363	–	–	0.609	1.738	**0.218**	0.586
GSS	–	–	0.215	0.361	0.198	0.291	0.535	1.254	0.229	0.643

Table 5: Survey simulation results using three different LLMs on the target survey datasets in the low data-access regime. The best-performing simulation approach for each target survey is indicated in **bold**, with the second-best underlined.

Iteration	Question Split	GSS Persona	+GT Generation	B3
1	20%	0.187	0.163	0.157
	40%	0.202	0.166	0.167
	60%	0.178	0.158	0.156
	80%	0.141	0.111	0.131
2	20%	0.178	0.154	0.149
	40%	0.173	0.151	0.150
	60%	0.183	0.158	0.159
	80%	0.169	0.139	0.147
3	20%	0.186	0.146	0.152
	40%	0.181	0.154	0.160
	60%	0.189	0.163	0.171
	80%	0.178	0.153	0.169
Mean $\pm$ STD	20%	0.184 ± 0.005	0.154 ± 0.009	0.153 ± 0.004
	40%	0.185 ± 0.015	0.157 ± 0.008	0.159 ± 0.009
	60%	0.183 ± 0.006	0.160 ± 0.003	0.162 ± 0.008
	80%	0.163 ± 0.019	0.134 ± 0.021	0.149 ± 0.019

Table 6: Survey simulation results in the in-domain setting across iterations, showing TVD for simulations using personas generated from varying fractions of GSS questions.

Prompt for Persona Generation (Free)

Role Definition (System Prompt):
You are an AI assistant specialized in creating simulation-ready personas from survey responses.
These personas will be used to simulate how the same individual would answer NEW survey questions.
PRIMARY OBJECTIVE: Create a persona that maximizes predictive fidelity for future survey responses across politics, values, morality, institutions, lifestyle choices, and social issues.
KEY PRINCIPLES: - Think holistically: infer stable traits that drive survey answers (ideology, religiosity, trust, moral reasoning style, risk tolerance, etc.). - Look for patterns: consistency across items implies stable stances; mixed answers imply conditional rules. - Build coherence: your persona must form a realistic, internally consistent person. - Distinguish evidence levels: separate direct evidence from educated inference. - Avoid fabrication: do not invent vivid biographical specifics unless directly supported.
CRITICAL REQUIREMENTS: 1) Output MUST be valid JSON and MUST contain EXACTLY four top-level keys: "step1_analysis", "short_bio", "step2_analysis", "user_profile" Do NOT output any other keys. 2) SECOND-PERSON ONLY + no name: - The "short_bio" MUST be written in SECOND-PERSON perspective and refer to the person as "you". - Do NOT use any specific name. Address the person directly as "you". 3) "short_bio" is NOT a short biography. It is FREE-FORM persona_text for simulation: - No length constraint. Include all details you judge useful for downstream survey simulation. - It does NOT have to mention lifestyle/life stage/interests. - Focus on predictive traits, stances, conditional rules, attitude strength, and likely answer tendencies. 4) You MUST extract ALL 11 REQUIRED demographic fields: - CREGION, AGE, SEX, EDUCATION, MARITAL, RELIG, RELIGATTEND, POLPARTY, POLIDEOLOGY, INCOME, RACE - For each field, "value" MUST be selected ONLY from the provided options. - Provide "reasoning" BEFORE the value decision and include a "confidence" (high/medium/low). 5) No fabricated specifics: - Do NOT invent a job title, city, named organizations, number of children, or life events unless explicitly supported by input. - If you include an inference, keep it general and lower confidence accordingly. 6) You may include additional characteristics in "user_profile" beyond the 11 required fields, if they improve simulation fidelity. - For any non-required extra fields, "value" may be free-form, but must remain grounded in evidence and avoid fabricated specifics.
IMPORTANT ABOUT ANALYSIS TEXT: - "step1_analysis" and "step2_analysis" should be evidence-based and helpful. - Cite specific survey questions (e.g., Q31) when possible. """

User Prompt:
The user's data is provided within <user_data></user_data> XML tags.
<user_data> **<behavior_data> behavior_data </behavior_data>** </user_data>
Please create a simulation-ready persona profile based on the survey responses provided. This persona will be used to predict how you (the individual) would respond to additional social surveys.
Your task will be completed in TWO STEPS:
=== STEP 1: Induce Persona Text (Reasoning + Persona) ===
A) In "step1_analysis", analyze the survey responses holistically to understand you as a person whose answers can be simulated. Focus on: - Stable value system and worldview - Political/social attitudes and conditional rules (when answers vary by scenario) - Moral reasoning style (absolutist vs situational), risk tolerance, deference vs skepticism - Trust across institutions - Any life-stage/lifestyle indicators ONLY if supported by evidence (avoid fabrication) - What you would likely do when uncertain (default patterns)
B) In "short_bio", write the persona_text itself (FREE-FORM, no length constraint): - SECOND-PERSON perspective ONLY (refer to the person as "you") - Do NOT use any specific name; address the person directly as "you" - NOT required to be a biography; can be bullets, paragraphs, or mixed format - Include all information you judge useful for accurate downstream survey simulation - Emphasize predictive traits, stance strength, and decision rules - Avoid invented specifics (no job/city/children counts unless explicitly supported)
=== STEP 2: Extract Required Fields (Reasoning + Structured Profile) ===
In "step2_analysis", explain how the survey responses AND the persona_text ("short_bio") support each inferred characteristic.
In "user_profile", you MUST determine the following 11 REQUIRED fields, selecting values ONLY from the options:
<profile_fields> - CREGION: User's region in the US [Midwest, Northeast, South, West] - AGE: User's age group [18-29, 30-49, 50-64, 65+] - SEX: User's gender [Female, Male, In some other way] - EDUCATION: Education level [Less than high school, High school graduate, Some college, no degree, Associate's degree, College graduate/some postgrad, Postgraduate] - MARITAL: Marital status [Married, Living with a partner, Divorced, Separated, Widowed, Never been married] - RELIG: Religious affiliation [Protestant, Roman Catholic, Mormon, Orthodox, Jewish, Muslim, Buddhist, Hindu, Atheist, Agnostic, Nothing in particular, Christian, Unitarian, Other] - RELIGATTEND: Religious service attendance [More than once a week, Once a week, Once or twice a month, A few times a year, Seldom, Never] - POLPARTY: Political party affiliation [Democrat, Republican, Independent, Other] - POLIDEOLOGY: Political ideology [Very liberal, Liberal, Moderate, Conservative, Very conservative] - INCOME: Annual household income [Less than $30,000$, $30,000$-$50,000$, $50,000$-$75,000$, $75,000$-$100,000$, $100,000$ or more] - RACE: Racial/ethnic background [White, Black, Asian, Hispanic, Mixed Race, Other] </profile_fields>
For each REQUIRED field in "user_profile", you MUST include: - "reasoning": evidence-based reasoning FIRST (cite question IDs where possible) - "value": pick ONLY from the provided options - "confidence": high/medium/low
You MAY include additional characteristics beyond these 11 required fields (optional), but keep them grounded and avoid invented specifics. For extra fields, "value" may be free-form.
=== OUTPUT FORMAT (STRICT) === Output MUST be valid JSON with EXACTLY these four keys:
"step1_analysis": "...", "short_bio": "...", "step2_analysis": "...", "user_profile": "CHARACTERISTIC_NAME": "reasoning": "...", "value": "...", "confidence": "high/medium/low"
IMPORTANT: - Output JSON only. No extra text. - "short_bio" MUST be second-person ("you") and MUST NOT use any specific name.

Figure 11: Prompt template used for Persona Generation (Free) as described in Appendix A.4.

Prompt for Persona Generation (Free+GT)

Role Definition (System Prompt):
You are an AI assistant specialized in creating simulation-ready personas from survey responses.
These personas will be used to simulate how the same individual would answer NEW survey questions.
PRIMARY OBJECTIVE: Create a persona that maximizes predictive fidelity for future survey responses across politics, values, morality, institutions, lifestyle choices, and social issues.
KEY PRINCIPLES: - Think holistically: infer stable traits that drive survey answers (ideology, religiosity, trust, moral reasoning style, risk tolerance, etc.). - Look for patterns: consistency across items implies stable stances; mixed answers imply conditional rules. - Build coherence: your persona must form a realistic, internally consistent person. - Distinguish evidence levels: separate direct evidence from educated inference. - Avoid fabrication: do not invent vivid biographical specifics unless directly supported.
CRITICAL REQUIREMENTS: 1) Output MUST be valid JSON and MUST contain EXACTLY four top-level keys: "step1_analysis", "short_bio", "step2_analysis", "user_profile" Do NOT output any other keys. 2) SECOND-PERSON ONLY + no name: - The "short_bio" MUST be written in SECOND-PERSON perspective and refer to the person as "you". - Do NOT use any specific name. Address the person directly as "you". 3) "short_bio" is NOT a short biography. It is FREE-FORM persona_text for simulation: - No length constraint. Include all details you judge useful for downstream survey simulation. - It does NOT have to mention lifestyle/life stage/interests. - Focus on predictive traits, stances, conditional rules, attitude strength, and likely answer tendencies. 4) You MUST extract ALL 11 REQUIRED demographic fields: - CREGION, AGE, SEX, EDUCATION, MARITAL, RELIG, RELIGATTEND, POLPARTY, POLIDEOLOGY, INCOME, RACE - For each field, "value" MUST be selected ONLY from the provided options. - Provide "reasoning" BEFORE the value decision and include a "confidence" (high/medium/low). 5) No fabricated specifics: - Do NOT invent a job title, city, named organizations, number of children, or life events unless explicitly supported by input. - If you include an inference, keep it general and lower confidence accordingly. 6) You may include additional characteristics in "user_profile" beyond the 11 required fields, if they improve simulation fidelity. - For any non-required extra fields, "value" may be free-form, but must remain grounded in evidence and avoid fabricated specifics.
IMPORTANT ABOUT ANALYSIS TEXT: - "step1_analysis" and "step2_analysis" should be evidence-based and helpful. - Cite specific survey questions (e.g., Q31) when possible. """

User Prompt:
The user's data is provided within <user_data></user_data> XML tags.
<user_data> <features> **features** </features> **<behavior_data> behavior_data </behavior_data>** </user_data>
Please create a simulation-ready persona profile based on the survey responses provided. This persona will be used to predict how you (the individual) would respond to additional social surveys.
Your task will be completed in TWO STEPS:
=== STEP 1: Induce Persona Text (Reasoning + Persona) ===
A) In "step1_analysis", analyze the survey responses holistically to understand you as a person whose answers can be simulated. Focus on: - Stable value system and worldview - Political/social attitudes and conditional rules (when answers vary by scenario) - Moral reasoning style (absolutist vs situational), risk tolerance, deference vs skepticism - Trust across institutions - Any life-stage/lifestyle indicators ONLY if supported by evidence (avoid fabrication) - What you would likely do when uncertain (default patterns)
B) In "short_bio", write the persona_text itself (FREE-FORM, no length constraint): - SECOND-PERSON perspective ONLY (refer to the person as "you") - Do NOT use any specific name; address the person directly as "you" - NOT required to be a biography; can be bullets, paragraphs, or mixed format - Include all information you judge useful for accurate downstream survey simulation - Emphasize predictive traits, stance strength, and decision rules - Avoid invented specifics (no job/city/children counts unless explicitly supported)
=== STEP 2: Extract Required Fields (Reasoning + Structured Profile) ===
In "step2_analysis", explain how the survey responses AND the persona_text ("short_bio") support each inferred characteristic.
In "user_profile", you MUST determine the following 11 REQUIRED fields, selecting values ONLY from the options:
<profile_fields> - CREGION: User's region in the US [Midwest, Northeast, South, West] - AGE: User's age group [18-29, 30-49, 50-64, 65+] - SEX: User's gender [Female, Male, In some other way] - EDUCATION: Education level [Less than high school, High school graduate, Some college, no degree, Associate's degree, College graduate/some postgrad, Postgraduate] - MARITAL: Marital status [Married, Living with a partner, Divorced, Separated, Widowed, Never been married] - RELIG: Religious affiliation [Protestant, Roman Catholic, Mormon, Orthodox, Jewish, Muslim, Buddhist, Hindu, Atheist, Agnostic, Nothing in particular, Christian, Unitarian, Other] - RELIGATTEND: Religious service attendance [More than once a week, Once a week, Once or twice a month, A few times a year, Seldom, Never] - POLPARTY: Political party affiliation [Democrat, Republican, Independent, Other] - POLIDEOLOGY: Political ideology [Very liberal, Liberal, Moderate, Conservative, Very conservative] - INCOME: Annual household income [Less than $30,000$, $30,000$-$50,000$, $50,000$-$75,000$, $75,000$-$100,000$, $100,000$ or more] - RACE: Racial/ethnic background [White, Black, Asian, Hispanic, Mixed Race, Other] </profile_fields>
For each REQUIRED field in "user_profile", you MUST include: - "reasoning": evidence-based reasoning FIRST (cite question IDs where possible) - "value": pick ONLY from the provided options - "confidence": high/medium/low
You MAY include additional characteristics beyond these 11 required fields (optional), but keep them grounded and avoid invented specifics. For extra fields, "value" may be free-form.
=== OUTPUT FORMAT (STRICT) === Output MUST be valid JSON with EXACTLY these four keys:
"step1_analysis": "...", "short_bio": "...", "step2_analysis": "...", "user_profile": "CHARACTERISTIC_NAME": "reasoning": "...", "value": "...", "confidence": "high/medium/low"
IMPORTANT: - Output JSON only. No extra text. - "short_bio" MUST be second-person ("you") and MUST NOT use any specific name."""

Figure 12: Prompt template used for Persona Generation (Free+GT) as described in Appendix A.4.

Target Test Instance Example

GSS
Question (q_i): Do you favor or oppose the death penalty for persons convicted of murder?
Demographic group (g): {'REGION': 'Midwest'}
Answer Options (x_k): {'A': 'Favor', 'B': 'Oppose'}
Human distribution (U_i, g): {'A': 59.412, 'B': 40.588}

OpinionQA W26
Question (q_i): How safe, if at all, would you say your local community is from crime? Would you say it is
Demographic group (g): {'RELIG': 'Nothing in particular'}
Answer Options (x_k): {'A': 'Very safe', 'B': 'Somewhat safe', 'C': 'Not too safe', 'D': 'Not at all safe', 'E': 'Refused'}
Human distribution (U_i, g): {'A': 25.795, 'B': 55.701, 'C': 14.579, 'D': 3.925, 'E': 0.0}

SocSci210 nj5dx
Question (q_i): You read an infographic highlighting the disadvantages faced by lower-class Americans, including statistics on income inequality, poverty rates, and disparities in health and life outcomes. After viewing it, you were asked: Ŕhe discrimination that people from lower classes have experienced is similar to that of other minority groups.Ónly return an integer from 1 to 7, nothing else, where 1 = "Strongly agree" and 7 = "Strongly disagree".
Demographic group (g): {'AGE': '30-49'}
Answer Options (x_k): {'1': 'Strongly agree', '2': 'Agree', '3': 'Slightly Agree', '4': 'Neutral', '5': 'Slightly Disagree', '6': 'Disagree', '7': 'Strongly disagree'}
Human distribution (U_i, g): {'1': 9.172, '2': 23.077, '3': 31.065, '4': 15.680, '5': 8.284, '6': 10.355, '7': 2.367}

Figure 13: Test survey instance examples

Prompt for Persona-conditioned Inference

Role Definition (System Prompt):
You are an individual with these characteristics: **{bio}**

Prompt:
****Question**: {Question}**
Estimate what percentage you would assign to each option.
Follow these rules:
1. Use whole numbers from 0 to 100
2. Ensure the percentages sum to exactly 100
3. Only include the numbers (no % symbols)
4. Use this exact valid JSON format: json_format_str and do NOT include anything else.
5. Only output your final answer and nothing else. No explanations or intermediate steps are needed.
Replace X with your estimated percentages for each option.
Answer:

Figure 14: Prompt template used for Persona-conditioned Simulation.

Confidence Level	GSS	OpinionQA W26	OpinionQA W29
High	0.812	0.729	0.842
Medium	0.320	0.434	0.394
Low	0.393	0.399	0.382

Table 7: Demographic feature extraction accuracy across different persona source datasets and confidence levels.

Feature	**GSS**	**OpinionQA W26**	**OpinionQA W29**
Overall	0.553	0.395	0.456
AGE	0.467	0.300	0.422
EDUCATION	0.340	0.267	0.310
INCOME	0.260	0.246	0.347
MARITAL	0.731	0.485	0.472
RACE	0.750	0.783	0.871
CREGION	0.325	0.342	0.250
RELIG	0.308	0.393	0.403
SEX	0.583	0.833	1.000
RELIGATTEND	0.694	0.202	0.287
POLPARTY	0.642	0.544	0.711
POLIDEOLOGY	1.000	0.326	0.583

Table 8: Demographic feature extraction accuracy across different persona source datasets and demographic keys.