When Can Digital Personas Reliably Approximate Human Survey Findings?

Mumin Jia
Department of Mathematics and Statistics
York University
Toronto, Ontario M3J 1P3
amyjia@yorku.ca

Yilin Chen
Department of Mathematics and Statistics
York University
Toronto, Ontario M3J 1P3
cathy929@yorku.ca

Divya Sharma
Department of Mathematics and Statistics
York University
Toronto, Ontario M3J 1P3
Department of Biostatistics
University Health Network
Toronto, ON M5G 2C4
divya03@yorku.ca

Jairo Diaz-Rodriguez
Department of Mathematics and Statistics
York University
Toronto, Ontario M3J 1P3
jdiazrod@yorku.ca

Abstract

Digital personas powered by Large Language Models (LLMs) are increasingly proposed as substitutes for human survey respondents, yet it remains unclear when they can reliably approximate human survey findings. We answer this question using the LISS panel, constructing personas from respondents' background variables and pre-2023 survey histories, then testing them against the same respondents' held-out post-cutoff answers. Across four persona architectures, three LLMs, and two prediction tasks, we assess performance at the question, respondent, distributional, equity, and clustering levels. Digital personas improve alignment with human response distributions, especially in domains tied to stable attributes and values, but remain limited for individual prediction and fail to recover multivariate respondent structure. Retrieval-augmented architectures provide the clearest gains, but performance depends more on human response structure than on model choice: personas perform best for low-variability questions and common respondent patterns, and worst for subjective, heterogeneous, or rare responses. Our results provide practical guidance on when digital personas could be appropriate for survey research and when human validation remains necessary.

1 Introduction

Large language models are increasingly being used to simulate human survey respondents [Argyle et al., 2023, Bisbee et al., 2024, Dominguez-Olmedo et al., 2024]. These "digital personas" promise faster, cheaper, and more flexible alternatives to conventional survey data collection, with potential uses in pre-testing survey instruments, exploring subgroup patterns, simulating hard-to-reach populations, and conducting early-stage social measurement before fielding studies with real participants. Yet their scientific value depends on more than whether their answers appear plausible: a synthetic respondent may reproduce population averages while failing to match the individuals it is meant to represent [Kaiser et al., 2025, Toubia et al., 2025], select majority responses while compressing minority or rare answer patterns, or preserve marginal distributions while distorting subgroup differences, respondent heterogeneity, and multivariate structure among survey items [Bisbee et al., 2024,

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

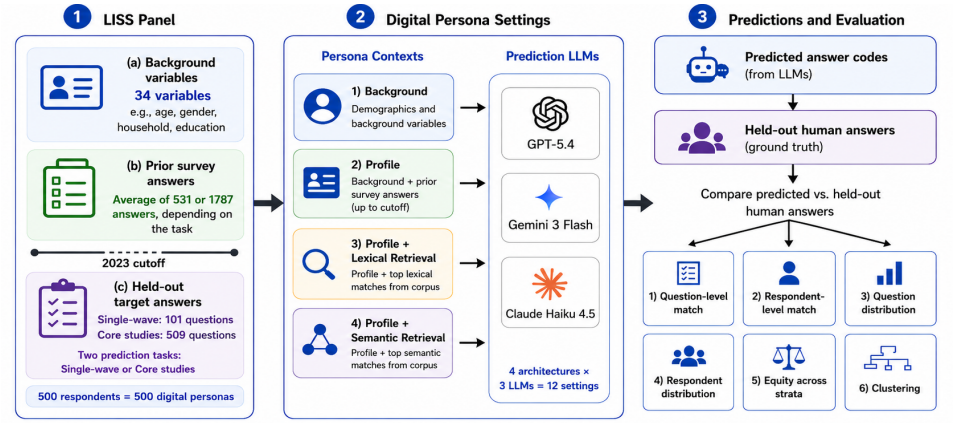


Figure 1: Overview of our digital persona evaluation framework.

Wang et al., 2025, Li et al., 2025a]. These failures matter because survey research is often used not only to estimate aggregate opinion, but also to study variation across people, groups, domains, and latent population structure.

This paper asks: *when can digital personas reliably approximate human survey findings?* We study this question in a ground-truth setting using the Longitudinal Internet Studies for the Social Sciences (LISS) panel, a probability-based longitudinal survey of Dutch households. For each respondent, we construct digital personas from information observed before a temporal cutoff and evaluate them against held-out answers from the same human respondent after that cutoff. This design tests whether persona-based simulations preserve real human responses, rather than merely generating demographically plausible or population-like answers.

Figure 1 provides an overview of our digital persona evaluation framework. Starting from the LISS panel, we use each respondent's background variables and pre-2023 survey histories to construct a digital persona, then ask that persona to predict the respondent's held-out post-cutoff survey answers. The framework varies both the information given to the persona and the model used for prediction: personas may receive only background variables, a structured profile, or a profile augmented with lexically or semantically retrieved prior answers, and predictions are generated with multiple LLM backbones. The resulting answer codes are then compared with the same respondents' true held-out answers. Rather than treating reliability as a single accuracy score, the framework evaluates whether digital personas preserve human survey findings across six complementary dimensions: question-level match, respondent-level match, question-level distributions, respondent-level response profiles, equity across demographic strata, and respondent clustering. This design allows us to ask not only whether digital personas can predict individual answers, but also whether they preserve the aggregate and structural patterns that survey researchers use for scientific inference. Overall, our study brings together data, modeling, and evaluation dimensions that prior work has typically studied separately. Table 1 shows that this combination has not been examined in a single empirical design.

Summary of findings. We provide an exploratory analysis of each reliability dimension in Section 5, followed by a confirmatory and explanatory modeling analysis in Section 6. Together, these analyses provide practical guidance on when digital personas are likely to be useful in survey research. First, digital personas align most closely with human response distributions in domains tied to stable attributes, values, and background information, such as family and household, politics and values, and religion and ethnicity. They perform worse in domains that depend on lived experience, self-assessment, or situational detail, such as social integration and leisure and personality. Second, retrieval-augmented persona contexts provide the clearest architectural benefit, while differences across the modern LLMs we evaluate are comparatively small. Third, performance appears broadly stable across gender, age group, and household stage, suggesting no large systematic demographic disparities. Fourth, reliability varies more strongly across the response space: personas are most accurate for low-variability questions and common respondent patterns, and least accurate for high-variability questions and rare respondent patterns. Question structure further shapes performance, with binary and small-answer-space questions easier to approximate than subjective questions or items likely to elicit diverse responses. Overall, digital personas are most useful for survey design

Table 1: Comparison of our study with related work. Green indicates that the study substantially includes the feature; yellow indicates a related but narrower or incomplete version; red indicates that the feature is absent or not central.

Study	Fixed-choice answers	Rep. stratified population sample	Same-person holdout	Temporal split	Real longitudinal panel	Prior survey history	Persona variants	Retrieval memory	Multi-LLM comparison	No-context baseline	Individual fidelity	Aggregate distribution	Equity / subgroups	Joint structure	Error drivers
Our study															
Argyle et al. [2023]															
Bisbee et al. [2024]															
Dominguez-Olmedo et al. [2024]															
Hwang et al. [2023]															
Park et al. [2026]															
Chen et al. [2025]															
Jiang et al. [2025]															
Li et al. [2025c]															
Toubia et al. [2025]															
Kaiser et al. [2025]															
Cao et al. [2025]															
Wang et al. [2025]															
Li et al. [2025a]															
Abdulhai et al. [2026]															
Morocho et al. [2026]															
Li et al. [2025b]															
Park et al. [2023]															
Moon et al. [2024]															
Hu and Collier [2024]															
Kim and Lee [2023]															
Suh et al. [2025]															
Kolluri et al. [2025]															
Hu et al. [2026]															
Rupprecht et al. [2025]															
Du et al. [2025]															
Wu et al. [2026]															
Chen et al. [2026]															
Hullman et al. [2026]															

and early-stage instrument testing, especially for identifying questions likely to yield stable aggregate patterns versus those requiring validation with human data.

Contributions. This work makes four contributions. First, we introduce a ground-truth evaluation design for digital personas in survey research using the LISS panel. Second, we propose a multi-dimensional reliability framework that evaluates personas at the question, respondent, distributional, equity, and clustering levels, thereby separating aggregate approximation from individual substitution. Third, we systematically compare four persona construction strategies and three LLM prediction models across two cross-domain prediction tasks. Fourth, we provide empirical evidence on the conditions under which digital personas appear more or less reliable, offering practical guidance for survey design and validation.

2 Position Relative to Prior Work

In Table 1, we position our study against prior work on synthetic survey respondents, digital personas, and persona reliability. The comparison emphasizes the combination of features needed for a strict test of digital-persona fidelity: objectively scored outcomes, a representative longitudinal panel, same-person temporal holdout prediction, respondent-specific prior survey histories, and evaluation beyond aggregate fit.

The main pattern is that prior work covers important but partial areas of this design space. Appendix E.1 defines the comparison criteria, and Appendix E provides the full related-work discussion and study-by-study comparison. Several of the closest studies are recent preprints that have not yet appeared in peer-reviewed venues, but the overall pattern is consistent: no single study combines the full set of data, persona-construction, and evaluation elements that we examine here.

3 Experiments

Our experiments study digital personas as conditional substitutes for real human participants in held-out survey prediction. Rather than asking whether LLMs can produce plausible survey answers in isolation, we construct tasks in which digital personas must predict the actual responses of real respondents from the LISS panel. We vary the available respondent history, the form of persona representation, the retrieval strategy, and the prediction model. This design provides a controlled setting for examining when digital personas have enough respondent-specific evidence to approximate real human answers, and when they do not.

3.1 Setup

Dataset. We use the Longitudinal Internet Studies for the Social Sciences (LISS) panel, a probability-based online panel of Dutch households administered by Centerdata [Mulder and Das, 2024]. The LISS panel is well suited to our setting because it repeatedly surveys the same individuals across multiple domains and waves, and also provides background variables describing respondents' sociodemographic characteristics. This structure allows us to construct a held-out prediction task: earlier responses and background information are used to build a digital persona, while later responses from the same respondent serve as ground truth.

Studies. The panel contains two broad types of questionnaires. *Core studies* are repeated longitudinal modules covering relatively stable domains such as personality, health, family, work, religion, politics, and social values. *Single-wave studies* are one-off surveys that often focus on topical or domain-specific questions. This distinction allows us to test generalization in both directions: from stable core histories to topical single-wave targets, and from broader single-wave histories to structured core-study targets.

Respondents. Using 2023 as the temporal cutoff, we retain respondents with a non-empty background file, non-empty prior-answer history, and at least one evaluable target answer. This yields 6,276 respondents with valid records. Each respondent is represented by an anonymized numeric identifier and has 34 sociodemographic background variables, including gender, age, household composition, urbanicity, income bracket, and related attributes.

Questions and responses. We restrict the analysis to closed-ended responses with finite answer spaces, including ordinal, true/false, nominal, and binary questions. We keep only standalone questions that can be answered from the question text alone, excluding items that require external context or refer to other survey items. The question metadata contain 2,923 unique standalone entries and specify the question text, variable name, response type, and allowed answer categories.

Prior and target answers. For each respondent, we split survey answers using 2023 as the temporal cutoff: pre-cutoff answers are used as prior evidence for persona construction, and held-out answers at or after the cutoff are used for evaluation. We define two prediction tasks:

- ***Single-wave prediction:*** prior evidence consists of the respondent's most recent pre-cutoff core-study answers, and targets are single-wave answers from 2023–2024.
- ***Core prediction:*** prior evidence consists of the respondent's full pre-cutoff single-wave answer history, and targets are core-study answers from 2023.

This design evaluates each digital persona against unseen responses from the same human respondent.

3.2 Respondent Sampling

For each task, we sample 500 eligible respondents from the LISS panel using stratified proportional sampling. Respondents are eligible if they have non-empty background variables, prior-answer history for the relevant input scope, and at least one evaluable target answer. Strata are defined by gender, age group, and household stage; strata with fewer than five eligible respondents are merged with the nearest compatible stratum. Within each stratum, we prioritize respondents with greater answer coverage using

$$s_i = z(n_i^{\text{prior}}) + z(n_i^{\text{target}}),$$

where n_i^{prior} and n_i^{target} denote the number of available prior and target answers for respondent i. The resulting samples preserve demographic variation while maintaining sufficient answer coverage; their composition is reported in Table 6 in Appendix. Detailed design is available in Appendix D. Appendix Figure 14 summarizes respondent-level answer coverage in the prior and target partitions used for evaluation after sampling.

3.3 Digital Personas

We define a digital persona as the prompt context used by an LLM to answer held-out survey questions for a specific respondent. For each prediction call, the LLM receives a block of target questions and a respondent-specific context, and returns predicted answer codes constrained to the admissible response options. We evaluate three prediction LLMs, GPT 5.4, Gemini 3 Flash, and Claude Haiku 4.5, crossed with four persona architectures, yielding 12 digital persona settings. The architecture determines what respondent information is included in the prompt context: background variables only, a structured profile, a profile with lexical retrieval, or a profile with semantic retrieval. We describe the architectures below, with full details provided in Appendix B and our GitHub repository.

Background. The background persona context contains only the respondent's 34 background variables and no prior survey answers. This is the minimal respondent-specific prompt condition and measures how much the prediction LLM can infer from sociodemographic information alone.

Profile. The profile persona context contains the respondent's background variables and a structured natural-language summary of their prior answers. The profile is generated once using GPT-5.4. The raw prior-answer rows are not included at prediction time. This condition tests whether a compact summary of survey history preserves useful respondent-specific information.

Profile + lexical retrieval. This persona context augments the structured profile with retrieved prior answers. For each prediction sub-batch, we append the K prior answered rows most lexically related to the current target questions. This condition tests whether surface-level similarity between prior and target survey items provides useful respondent-specific evidence.

Profile + semantic retrieval. This persona context uses the same structure as lexical retrieval, but replaces token overlap with embedding-based retrieval. Prior answered rows and target questions are embedded offline. For each prediction sub-batch, each prior row is scored by summing its cosine similarities to all target-question embeddings in the sub-batch. The top-K prior rows are appended to the prompt. This condition tests whether semantic similarity retrieves useful respondent memories when related survey items are phrased differently.

Baseline. As a baseline, we remove respondent-specific information entirely. The prediction LLM receives only the target question text, variable name, representation type, admissible answer categories, and answer constraints. Each target question is queried 500 times, matching the number of sampled respondents, so that the generated answer distribution can be compared with the empirical human distribution without using respondent information. This baseline separates gains from persona conditioning from performance driven by general LLM priors or majority-response tendencies. Appendix B.7 clarifies the baseline design choices.

4 Evaluation Framework

We evaluate digital persona reliability from six complementary dimensions: exact match at the question and respondent levels, distributional alignment at the question and respondent levels, equity across demographic strata, and preservation of respondent clusters. Together, these analyses distinguish individual-level prediction from aggregate survey approximation and test whether persona errors are evenly distributed or concentrated in particular groups or response structures.

Question dimension. We evaluate exact-response prediction at the question level. For each target question, we compare the answers generated by digital personas with the corresponding human answers across respondents. Performance is measured using weighted F1-score. This angle



Figure 2: Aggregate reliability of digital persona settings across evaluation dimensions. Radar plots summarize performance for the single-wave and core-study prediction tasks across question-level match, respondent-level match, question-level distributional alignment, respondent-level distributional alignment, equity, and clustering. Higher values indicate better performance.

Table 2: Fidelity of digital personas on core predictions across five persona conditions (1 baseline + 4 personas) and three LLMs. Metrics measure Clustering ARI ↑, Respondent exact match rate ↑, Question weighted F1 score ↑, Question Distribution JSD ↓, Respondent Distribution MMD ↓, and equity mean absolute deviation of DPI ↓. Core studies involve richer question entries and respondents, yielding higher persona-fidelity headroom compared to single-wave predictions. Baseline is GPT-5.4 only. Bold denotes values statistically tied with the best-performing setting under overlap of the reported uncertainty intervals.

		Baseline	Background	Profile	Profile+Lexical	Profile+Semantic
Clustering ARI (↑)	**GPT**	0.011 (±0.011)	0.000 (±0.002)	0.000 (±0.004)	0.000 (±0.037)	0.000 (±0.020)
	Claude	—	0.009 (±0.005)	0.011 (±0.041)	**0.133 (±0.068)**	0.000 (±0.060)
	Gemini	—	0.000 (±0.005)	0.000 (±0.006)	0.000 (±0.007)	0.000 (±0.002)
Resp. Match rate (↑)	**GPT**	0.478 (±0.002)	0.507 (±0.002)	0.521 (±0.002)	0.529 (±0.002)	0.531 (±0.002)
	Claude	—	0.486 (±0.002)	0.497 (±0.002)	0.510 (±0.002)	0.510 (±0.002)
	Gemini	—	0.515 (±0.002)	0.528 (±0.002)	**0.536 (±0.002)**	**0.534 (±0.002)**
Question F1 Score (↑)	**GPT**	0.306 (±0.003)	0.392 (±0.004)	0.419 (±0.003)	0.434 (±0.003)	**0.436 (±0.004)**
	Claude	—	0.367 (±0.003)	0.399 (±0.003)	0.412 (±0.004)	0.415 (±0.003)
	Gemini	—	0.402 (±0.004)	0.423 (±0.004)	**0.444 (±0.004)**	**0.444 (±0.004)**
Ques. Dist. JSD (↓)	**GPT**	0.530 (±0.004)	0.375 (±0.005)	0.338 (±0.005)	0.315 (±0.005)	**0.306 (±0.005)**
	Claude	—	0.409 (±0.004)	0.370 (±0.004)	0.345 (±0.004)	0.338 (±0.004)
	Gemini	—	0.365 (±0.005)	0.339 (±0.005)	**0.305 (±0.005)**	**0.301 (±0.005)**
Resp. Dist. MMD (↓)	**GPT**	0.343 (±0.002)	0.215 (±0.002)	0.181 (±0.002)	0.161 (±0.002)	**0.155 (±0.002)**
	Claude	—	0.241 (±0.002)	0.202 (±0.002)	0.176 (±0.002)	0.179 (±0.002)
	Gemini	—	0.204 (±0.002)	0.184 (±0.002)	0.165 (±0.002)	**0.157 (±0.002)**
Equity DPI (↓)	**GPT**	0.017 (±0.003)	0.015 (±0.002)	0.013 (±0.002)	**0.012 (±0.002)**	**0.011 (±0.002)**
	Claude	—	0.022 (±0.002)	0.016 (±0.002)	0.014 (±0.002)	**0.011 (±0.002)**
	Gemini	—	**0.007 (±0.002)**	**0.009 (±0.003)**	**0.009 (±0.002)**	**0.009 (±0.002)**

helps identify which survey questions are reliably approximated and which item characteristics are associated with higher or lower performance.

Respondent dimension. We evaluate prediction accuracy at the respondent level by measuring how well each digital persona reproduces the corresponding human respondent's answers across target questions. For each respondent, we compute exact match rate over the questions they answered. This angle helps identify which respondents are easier or harder to approximate and allows us to study whether fidelity varies with respondent characteristics.

Question-Distribution dimension. We evaluate whether digital personas preserve population-level response distributions. For each target question, we compare digital persona answer distributions with

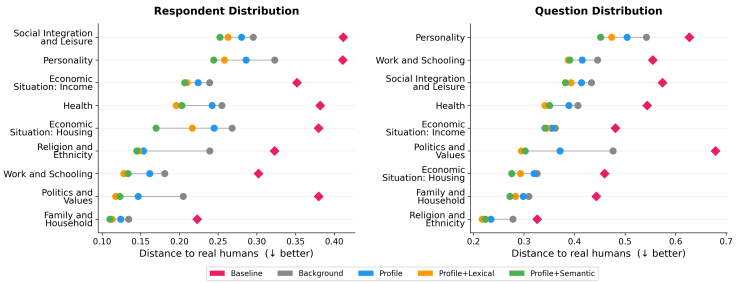


Figure 3: Distributional distance to human responses by study domain. Lower values indicate closer alignment with the empirical human distribution. The left panel compares respondent-level response distributions, while the right panel compares question-level response distributions.

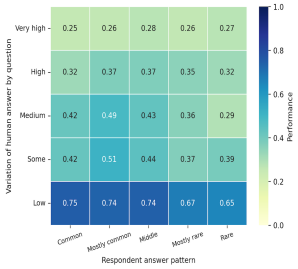


Figure 4: Digital persona performance by question-level answer variability and respondent answer pattern.

empirical human answer distributions for the same question. We measure this distance using Jensen–Shannon divergence (JSD) (C.1.1) [Lin, 1991]. Lower JSD indicates closer alignment between synthetic and human response distributions. This captures whether personas preserve aggregate survey findings even when individual-level exact matches are imperfect.

Respondent-distribution dimension. We evaluate whether digital personas preserve each respondent's overall response profile across questions. Because each respondent is represented by a multivariate distribution over answers across many survey items, we use maximum mean discrepancy (MMD) to compare the digital persona response distribution with the human response distribution (C.1.1) [Gretton et al., 2012]. This angle captures whether the persona preserves the respondent's overall pattern of answer usage, such as agreement, disagreement, extremity, or category preference.

Equity dimension. We evaluate whether persona reliability varies across the demographic strata used for respondent sampling. Specifically, we compare performance across gender, age group, and household-stage categories using the mean absolute deviation of the Demographic Parity Index (DPI) across strata (C.1.1). This analysis measures whether prediction errors are distributed evenly across groups or concentrated within particular strata. It is important because aggregate performance may appear stable even when reliability differs systematically across groups.

Clustering dimension. We evaluate whether digital persona responses preserve higher-level respondent structure. We cluster respondents using their observed human answers and separately cluster digital personas using their generated answers, then compare the resulting partitions using Adjusted Rand Index (ARI) (C.1.1) [Hubert and Arabie, 1985]. Higher ARI indicates stronger agreement between human-based and persona-based respondent clusters. This angle helps test whether personas preserve multivariate population structure relevant for segmentation, typology discovery, and subgroup analysis.

5 Results

We generate held-out responses for all digital persona settings and the no-context baseline in both 2023 prediction tasks: core-study prediction and single-wave prediction. We now discuss the main findings from these predictions.

5.1 Overview of Persona Reliability

Figure 2 summarizes performance across the six evaluation dimensions for both prediction tasks. All metrics are scaled from 0 to 1, with higher values indicating better performance. Overall, the results show a consistent pattern. Equity scores are stable across settings, suggesting no large disparities across the demographic strata used in our sample. Clustering performance is weak across all settings, indicating that persona-generated responses do not reliably recover the respondent structure induced by human answers. Question-level and respondent-level exact-match performance improve over the baseline, but remain limited, suggesting that digital personas are not reliable substitutes for individual

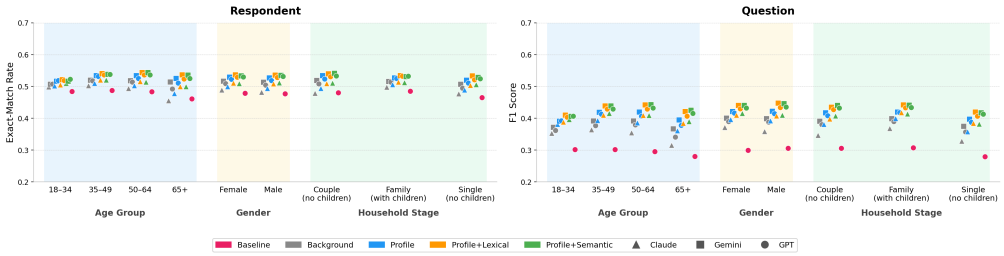


Figure 5: Exact-match performance across demographic strata at the respondent level (left, measured by exact match rate) and question level (right, measured by weighted F1 score).

respondents. In contrast, the clearest advantage appears in distributional alignment: persona-based settings more closely approximate human response distributions and consistently outperform the no-context baseline. Table 2 reports the corresponding core-study results across all architectures and LLMs, including bootstrapped confidence intervals, while the corresponding single-wave results are provided in Table 5 in the appendix. The results show that retrieval-augmented digital personas are consistently among the best-performing settings across nearly all metrics, with differences supported by the reported confidence intervals. Across prediction LLMs, gains are small compared with the gains from persona architecture and retrieval. We therefore interpret model choice as secondary.

5.2 Distributional Alignment

Figure 3 compares the distance between human response distributions and the distributions produced by each setting for the core-study prediction task. Across both respondent-level and question-level views, the baseline is consistently farther from humans than the digital personas, indicating that baseline predictions may exploit majority-answer patterns but do not reproduce the empirical distribution of human responses. Digital personas substantially reduce distributional distance across most domains, with retrieval variants consistently closest to the human distributions.

Digital personas align most closely with human distributions in family and household, politics and values, and religion and ethnicity. These domains may be easier to approximate because they are closely tied to information available in the persona context. Family and household responses are often reflected in background variables such as age, marital status, children, and household composition. Politics, values, religion, and ethnicity tend to be more stable over time and may be captured by prior attitude- and identity-related survey answers. In contrast, social integration and leisure and personality show larger distributional distances. These domains depend more on individual lived experience and self-assessment, rather than broad demographic or ideological signals. As a result, they are harder for digital personas to infer unless the prior history contains directly related evidence.

5.3 Equity and Conditional Exact-Match Performance

Figure 5 reports performance across the demographic strata used in sampling. Respondent-level performance (left) and question-level performance (right) are broadly stable across gender, age group, and household stage, with no large systematic disparities. This suggests that persona performance is not driven by a single demographic group and is relatively balanced across the sampled strata. Across architectures, retrieval-augmented settings perform slightly best overall. Across LLMs, Gemini performs slightly better than GPT, while Claude performs weakest.

Figure 4 shows a different source of variation. Performance changes substantially when conditioned on question-level human answer variability and respondent answer commonness. Accuracy is highest for low-variability questions and common respondent patterns, and lowest for high-variability questions and rare respondent patterns. Thus, while performance appears demographically balanced, persona reliability is still uneven: errors concentrate where human responses are more heterogeneous and less common. Details on the construction of the heatmap are provided in Appendix C.1.2.

5.4 Clustering and Multivariate Respondent Structure

Figure 9 in the appendix evaluates whether digital persona responses recover the respondent clusters obtained from human answers. Across settings, clustering agreement is generally low, indicating

Table 3: Core study prediction performance across feature layers ($n = 132{,}887$ observations).

	Behavioral		Contextual		Structural	
Model	ACC	AUC	ACC	AUC	ACC	AUC
Logistic Regression	0.706 ($\pm$0.001)	0.775	0.658 ($\pm$0.002)	0.722	0.632 ($\pm$0.002)	0.700
MEGLM	0.705 ($\pm$0.001)	0.765	0.653 ($\pm$0.001)	0.719	0.643 ($\pm$0.001)	0.700
Decision Tree	0.699 ($\pm$0.001)	0.802	0.681 ($\pm$0.002)	0.738	0.672 ($\pm$0.002)	0.732
Random Forest	0.727 ($\pm$0.002)	**0.819**	0.695 ($\pm$0.002)	0.758	0.681 ($\pm$0.002)	**0.744**
XGBoost	**0.757** ($\pm$0.002)	0.817	**0.705** ($\pm$0.002)	**0.769**	**0.685** ($\pm$0.002)	0.743

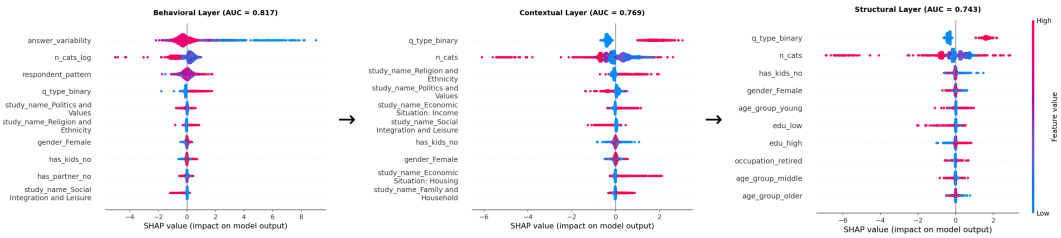


Figure 6: Top 10 predictors of digital persona accuracy from XGBoost models across three explanatory feature layers. Each point is a respondent-question prediction; the x-axis shows the SHAP value, where positive values increase the model's predicted probability of a correct persona answer and negative values decrease it. Color indicates the feature value, from low to high. The behavioral layer includes empirical response-distribution features, the contextual layer removes these distribution-derived variables, and the structural layer further focuses on observable respondent and question features.

that synthetic response matrices do not preserve the same multivariate respondent structure as the human data. This provides an important caution: even when digital personas improve question- or respondent-level distributional alignment, they may still fail to preserve higher-order dependence patterns among answers. More details and additional analysis available in Appendix C.1.3.

6 What Predicts Whether a Digital Persona Matches a Human Response?

The results above suggest that digital personas are more reliable for distributional approximation than for exact individual prediction, with performance varying across study domains and regions of the response space. We use respondent-question-level modeling as a confirmatory and explanatory analysis, testing whether these patterns remain predictive of individual-level accuracy when question, study, and respondent features are considered jointly. For this analysis, we focus on the GPT-5.4 profile with lexical retrieval setting. Each observation is one predicted answer for one respondent on one target question, with a binary outcome indicating whether the digital persona exactly matches the human answer. We compare logistic regression, mixed-effects logistic regression, decision trees, random forests, and XGBoost across three feature layers: behavioral, contextual, and structural. Table 3 reports the predictive performance for core-study prediction. Tree-based models perform best overall, with XGBoost achieving the highest accuracy across all three feature layers while remaining among the strongest models in AUC. We therefore use XGBoost for the feature-attribution analysis in Figure 6, which reports the top ten SHAP features from the best XGBoost models for core prediction. Features are ordered by importance, so variables appearing near the top contribute most to the model's predictions. Additional modeling details, bootstrapped confidence intervals, and corresponding results for single-wave prediction are provided in Appendix C.2.

Behavioral layer. The behavioral layer includes empirical response-structure features from Figure 4, including question-level answer variability and respondent answer pattern. These features dominate the model. Answer variability is the strongest predictor of persona accuracy, followed by the number of answer categories and respondent answer pattern. This supports the result from Figure 4: personas are most accurate when human answers are concentrated and respondent patterns are common, and least accurate when questions elicit heterogeneous or rare responses. Question format also matters: larger answer spaces reduce performance, while binary questions increase it.

Contextual layer. The contextual layer removes behavioral predictors but retains broader survey context, including the substantive domain and format of the target question. This approximates a prospective setting in which empirical response distributions are unavailable, but the general meaning of the question is known. In this layer, question structure remains highly predictive: larger answer spaces reduce performance, while binary questions increase it. Domain context is also important, consistent with Figure 3: personas perform better in domains tied to stable attributes, values, and background information, and worse in domains that depend more on lived experience, self-assessment, or situational detail.

Structural layer. The structural layer further removes study identifiers, leaving question form and respondent attributes. Question type and number of answer categories remain the strongest predictors, while respondent characteristics have smaller effects. This qualifies the demographic results in Figure 5: performance is broadly balanced across sampled strata, but demographics still explain some residual variation once stronger behavioral and contextual predictors are removed.

Overall, this analysis ties the empirical results back to the central question of the paper: when can digital personas approximate human survey findings? By identifying the conditions under which persona predictions are more or less reliable, it provides a guide for practical use and motivates the conclusion below.

7 Conclusion

Our findings provide a practical guideline for using digital personas in survey research. Practitioners should use them most cautiously for individual-level prediction and most confidently for aggregate distributional approximation, especially when questions have small answer spaces and human responses are likely to be concentrated. Digital personas appear more reliable in domains tied to stable attributes, values, and background information, such as family and household, politics and values, and religion and ethnicity, and less reliable in domains that depend on lived experience, self-assessment, or situational detail. Their performance also appears broadly balanced across the demographic strata we study, suggesting that errors are not driven mainly by gender, age group, or household stage. Across the modern LLMs we evaluate, performance differences are relatively small, while retrieval-augmented persona contexts provide the clearest architectural benefit. Notably, the limited gains from more sophisticated persona architectures are themselves informative: they suggest that the primary bottleneck for digital persona reliability is not the richness of the retrieval mechanism or the choice of model, but the structure of the response space itself. However, reliability varies strongly with the structure of the response task: personas are less appropriate when the target requires recovering rare, heterogeneous, subjective, or highly respondent-specific answers. In practice, persona reliability should therefore be assessed against the intended inferential use and the structure of the survey task. The more variable, subjective, or idiosyncratic the expected human responses are, the more important it is to validate persona outputs against real human data.

Limitations. Our evaluation is conducted on a single longitudinal panel of Dutch households, and findings may not generalize to panels from other countries, languages, or cultural contexts. The framework is designed to be applicable to other longitudinal surveys with repeated measurements of the same respondents, but cross-panel validation remains important future work. The analysis is restricted to closed-ended questions with finite answer spaces, leaving open how digital personas perform on open-ended or scale-based items. Our respondent sample is capped at 500 per task due to computational constraints, and results may differ with larger or more diverse samples. Finally, persona profiles are generated using GPT-5.4. Future work should examine whether profile quality varies meaningfully across profile-generating models.

Acknowledgments and Disclosure of Funding

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) under grants DGECR-2022-04531 and RGPIN-2024-05548. The authors thank Jianan Lei, the anonymous reviewers, and the session chair for their valuable feedback and insightful suggestions, which greatly improved the quality of this work.

References

Marwa Abdulhai, Ryan Cheng, Donovan Clay, Tim Althoff, Sergey Levine, and Natasha Jaques. Consistently simulating human personas with multi-turn reinforcement learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=A0T3piHiis.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023. doi: 10.1017/pan.2023.2.

James Bisbee, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M. Larson. Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, 32 (4):401–416, 2024. doi: 10.1017/pan.2024.5.

Yong Cao, Haijiang Liu, Arnav Arora, Isabelle Augenstein, Paul Röttger, and Daniel Hershcovich. Specializing large language models to simulate survey response distributions for global populations. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *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, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.162. URL https://aclanthology.org/2025.naacl-long.162/.

Jiawei Chen, Ruoxi Xu, Boxi Cao, Ruotong Pan, Yunfei Zhang, Yifei Hu, Yong Du, Tingting Gao, Yaojie Lu, Yingfei Sun, Xianpei Han, Le Sun, Xiangyu Wu, and Hongyu Lin. Towards real-world human behavior simulation: Benchmarking large language models on long-horizon, cross-scenario, heterogeneous behavior traces, 2026. URL https://arxiv.org/abs/2604.08362.

Sihan Chen, John P. Lalor, Yi Yang, and Ahmed Abbasi. PersonaTwin: A multi-tier prompt conditioning framework for generating and evaluating personalized digital twins. In Ofir Arviv, Miruna Clinciu, Kaustubh Dhole, Rotem Dror, Sebastian Gehrmann, Eliya Habba, Itay Itzhak, Simon Mille, Yotam Perlitz, Enrico Santus, João Sedoc, Michal Shmueli Scheuer, Gabriel Stanovsky, and Oyvind Tafjord, editors, *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 774–788, Vienna, Austria and virtual meeting, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-261-9. URL https://aclanthology.org/2025.gem-1.66/.

Ricardo Dominguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dünner. Questioning the survey responses of large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=0o7dlLgqQX.

Bangde Du, Minghao Guo, Songming He, Ziyi Ye, Xi Zhu, Weihang Su, Shuqi Zhu, Yujia Zhou, Yongfeng Zhang, Qingyao Ai, and Yiqun Liu. TwinVoice: A multi-dimensional benchmark towards digital twins via LLM persona simulation, 2025. URL https://arxiv.org/abs/2510.25536.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL http://jmlr.org/papers/v13/gretton12a.html.

Tiancheng Hu and Nigel Collier. Quantifying the persona effect in LLM simulations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 10289–10307, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.554. URL https://aclanthology.org/2024.acl-long.554/.

Tiancheng Hu, Joachim Baumann, Lorenzo Lupo, Dirk Hovy, Nigel Collier, and Paul Röttger. SimBench: Benchmarking the ability of large language models to simulate human behaviors. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=PL51SpN6ZJ. ICLR 2026 poster.

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985. doi: 10.1007/BF01908075. URL https://doi.org/10.1007/BF01908075.

Jessica Hullman, David Broska, Huaman Sun, and Aaron Shaw. This human study did not involve human subjects: Validating LLM simulations as behavioral evidence, 2026. URL https://arxiv.org/abs/2602.15785.

EunJeong Hwang, Bodhisattwa Majumder, and Niket Tandon. Aligning language models to user opinions. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5906–5919, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.393. URL https://aclanthology.org/2023.findings-emnlp.393/.

Bowen Jiang, Zhuoqun Hao, Young Min Cho, Bryan Li, Yuan Yuan, Sihao Chen, Lyle Ungar, Camillo Jose Taylor, and Dan Roth. Know me, respond to me: Benchmarking LLMs for dynamic user profiling and personalized responses at scale. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=6ox8XZGOqP.

Carolin Kaiser, Jakob Kaiser, Vladimir Manewitsch, Lea Rau, and Rene Schallner. Simulating human opinions with large language models: Opportunities and challenges for personalized survey data modeling. In *Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization*, UMAP Adjunct '25, page 82–86, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713996. doi: 10.1145/3708319.3733685. URL https://doi.org/10.1145/3708319.3733685.

Junsol Kim and Byungkyu Lee. AI-augmented surveys: Leveraging large language models and surveys for opinion prediction, 2023. URL https://arxiv.org/abs/2305.09620.

Akaash Kolluri, Shengguang Wu, Joon Sung Park, and Michael S. Bernstein. Finetuning LLMs for human behavior prediction in social science experiments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.emnlp-main.1530/.

Ang Li, Haozhe Chen, Hongseok Namkoong, and Tianyi Peng. LLM generated persona is a promise with a catch. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems Position Paper Track*, 2025a. URL https://openreview.net/forum?id=qh9eGtMG4H.

Chance Jiajie Li, Zhenze Mo, Yuhan Tang, Ao Qu, Jiayi Wu, Kaiya Ivy Zhao, Yulu Gan, Jie Fan, Jiangbo Yu, Jinhua Zhao, Paul Pu Liang, Luis Alberto Alonso Pastor, and Kent Larson. Hugagent: Evaluating LLMs in simulating individual-level human reasoning on open-ended tasks. In *NeurIPS 2025 Workshop on Bridging Language, Agent, and World Models for Reasoning and Planning*, 2025b. URL https://openreview.net/forum?id=bxR1cSF4dt.

Rui Li, Heming Xia, Xinfeng Yuan, Qingxiu Dong, Lei Sha, Wenjie Li, and Zhifang Sui. How far are LLMs from being our digital twins? a benchmark for persona-based behavior chain simulation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 15738–15763, Vienna, Austria, July 2025c. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.813. URL https://aclanthology.org/2025.findings-acl.813/.

J. Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991. doi: 10.1109/18.61115.

Suhong Moon, Marwa Abdulhai, Minwoo Kang, Joseph Suh, Widyadewi Soedarmadji, Eran Kohen Behar, and David M. Chan. 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, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1110. URL https://aclanthology.org/2024.emnlp-main.1110/.

Erika Elizabeth Taday Morocho, Lorenzo Cima, Tiziano Fagni, Marco Avvenuti, and Stefano Cresci. Assessing the reliability of persona-conditioned llms as synthetic survey respondents, 2026. URL https://arxiv.org/abs/2602.18462.

Joris Mulder and Marcel Das. Referencing and citing the LISS panel, August 2024. URL https://www.lissdata.nl/app/uploads/sites/4/2024/08/Reference_LISS_v6.1_20240828.pdf.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA, 2023. Association for Computing Machinery. doi: 10.1145/3586183.3606763. URL https://doi.org/10.1145/3586183.3606763.

Joon Sung Park, Carolyn Q. Zou, Jonne Kamphorst, Niles Egan, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Percy Liang, Robb Willer, and Michael S. Bernstein. Llm agents grounded in self-reports enable general-purpose simulation of individuals, 2026. URL https://arxiv.org/abs/2411.10109.

Jens Rupprecht, Leon Fröhling, Claudia Wagner, and Markus Strohmaier. German general social survey personas: A survey-derived persona prompt collection for population-aligned LLM studies, 2025. URL https://arxiv.org/abs/2511.21722.

Joseph Suh, Erfan Jahanparast, Suhong Moon, Minwoo Kang, and Serina Chang. 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*, pages 21147–21170. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.acl-long.1028/.

Olivier Toubia, George Z. Gui, Tianyi Peng, Daniel J. Merlau, Ang Li, and Haozhe Chen. Twin-2k-500: A data set for building digital twins of over 2,000 people based on their answers to over 500 questions. *Marketing Science*, 44(6):1446–1455, 2025. doi: 10.1287/mksc.2025.0262.

Angelina Wang, Jamie Morgenstern, and John P. Dickerson. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, 7(3):400–411, March 2025. ISSN 2522-5839. doi: 10.1038/s42256-025-00986-z. URL https://doi.org/10.1038/s42256-025-00986-z.

Shirley Wu, Evelyn Choi, Arpandeep Khatua, Zhanghan Wang, Joy He-Yueya, Tharindu Cyril Weerasooriya, Wei Wei, Diyi Yang, Jure Leskovec, and James Zou. HumanLM: Simulating users with state alignment beats response imitation, 2026. URL https://arxiv.org/abs/2603.03303.

Siyan Zhao, John Dang, and Aditya Grover. Group preference optimization: Few-shot alignment of large language models. In *International conference on learning representations*, volume 2024, pages 57965–57987, 2024.

Appendix

A Data Use and License Statement **15**

B Experimental Details **15**

B.1 Prediction Batching 15

B.2 Prompt Templates 15

B.2.1 Background Agent 16

B.2.2 Profile Agent: Profile Generation 16

B.2.3 Profile Agent: Prediction Prompt 17

B.2.4 Retrieval-Augmented Agents: Prediction Prompt 17

B.3 Lexical Retrieval 18

B.4 Semantic Retrieval 18

B.5 Answer Validation and Retry 18

B.6 Models 19

B.7 Clarification on Baselines 19

C Additional Experiments and Results **19**

C.1 Supplementary Results for Section 5 19

C.1.1 Detailed Results of All Dimensions 19

C.1.2 Additional Details on Figure 4 20

C.1.3 Clustering 21

C.2 Supplementary Experiments for Section 6 23

D Sampling Design **25**

E Positioning Relative to Prior Work (Expanded) **27**

E.1 Comparison criteria 30

F Computational Resources and Cost **31**

G Broader Impacts **31**

A Data Use and License Statement

This study uses data from the LISS panel. In accordance with the required acknowledgement, we state: "In this paper, we make use of data from the LISS panel (Longitudinal Internet studies for the Social Sciences) managed by the non-profit research institute Centerdata (Tilburg University, the Netherlands)."

The LISS panel data are not redistributed with this manuscript. Access to the data is governed by Centerdata's LISS Data Archive rules and user statement. Published datasets can be downloaded free of charge only for non-commercial scientific or policy-relevant purposes, subject to registration and acceptance of the LISS data user statement. Access is personal, and data users are not permitted to make copies of the data available to others.

Researchers who wish to access the data should request access directly through the LISS Data Archive at `https://www.lissdata.nl`. After creating an account and joining the LISS community, users can search for and download the relevant LISS panel datasets, subject to Centerdata's access conditions and user statement.

Nevertheless, all results reported in this manuscript are fully reproducible by users who obtain permission to access the relevant LISS panel data and use the accompanying replication materials provided by the authors.

All analyses, interpretations, and conclusions in this manuscript are the responsibility of the authors and do not necessarily reflect the views of Centerdata, Tilburg University, or the LISS panel.

B Experimental Details

We report complete implementation details for prediction batching, prompt construction, retrieval scoring, answer validation, and model configuration. Code and replication materials are available at `https://github.com/muminjia/digital-persona-suppl`.

B.1 Prediction Batching

Target questions are grouped by their originating study (`study_id` for single-wave targets; `project_number` for core targets) and partitioned into sub-batches. For each group s containing $N_g^{(s)}$ questions and a target sub-batch size $b = 20$, the number of sub-batches and the realized sub-batch size are

$$B^{(s)} = \left\lceil \frac{N_g^{(s)}}{b} \right\rceil, \qquad N_{\text{batch}}^{(s)} = \left\lceil \frac{N_g^{(s)}}{B^{(s)}} \right\rceil. \tag{1}$$

Each API call covers exactly one sub-batch. For the two retrieval-augmented agents, the top-K retrieval count is set dynamically to $K = N_{\text{batch}}^{(s)}$, coupling context size to the number of questions being predicted in that call.

B.2 Prompt Templates

All four agents share the same system prompt and output schema; they differ only in what respondent evidence appears in the user turn.

System prompt. Every API call uses the system message: *"You map user profile data to survey answers carefully and transparently."*

Output schema. Each prediction call must return a JSON object containing a `predictions` array, where each entry maps a `variable_name` to its `predicted_answer`. No additional keys, markdown, or prose are permitted.

B.2.1 Background Agent

The Background agent receives only the 34-variable background file as context; no prior survey answers are included. The user-turn prompt is:

```
You are simulating a specific person answering a survey.

Answer each question as this person would, using only the
respondent background data provided below.

Background data:
{user_context}

Instructions:
- Answer as the person.
- Use the background data to infer the most likely response.
- If the answer is uncertain, choose the response most
  consistent with the available evidence.
- Provide exactly one answer per question.

Return valid JSON only. No markdown, code fences, or extra text.

Questions:
{questions_text}
```

B.2.2 Profile Agent: Profile Generation

Before prediction, a separate LLM call condenses the respondent's prior answered rows into a structured persona profile. Profile generation uses GPT-5.4 and is executed once per respondent per input scope; the resulting text is cached on disk and reused by all subsequent prediction calls for that respondent, including those of the two retrieval-augmented agents. The profile-generation prompt, where `{history_text}` is the full serialization of all prior answered rows in the selected input scope, is:

```
You are analyzing a set of questions and answers from a single person.

Build a predictive persona profile that allows another model to
estimate how this person would respond to future survey questions
on any topic.

From the provided Q&A, extract stable behavioral patterns across:
- beliefs and worldview
- personality traits
- reasoning and decision-making style
- knowledge domains and expertise
- tolerance for uncertainty
- recurring assumptions and cognitive biases
- communication style and priorities

Do not summarize individual answers. Identify consistent rules
and tendencies that generalize across questions.

Structure your output under the following seven headings:
  1. Personality traits
  2. Reasoning style
  3. Knowledge profile
  4. Values and motivations
  5. Biases and heuristics
  6. Decision patterns
  7. Confidence patterns

Past Q&A:
{history_text}
```

The profile must contain all seven section headers and must not exceed 1,500 words. If either constraint is violated, the call is retried with the validation error appended to the prompt; generation is attempted up to three times before the run aborts. A cached profile is identified by user ID, dataset source, cutoff year, input scope, and profile model; if a matching file exists on disk, the generation call is skipped entirely.

B.2.3 Profile Agent: Prediction Prompt

The profile agent's prediction prompt replaces the raw prior-answer rows with the synthesized persona profile:

```
You are simulating a specific person answering a survey.

Answer each question as this person would, using:
  1. the behavioral profile below
  2. the respondent background data as supporting context

Background data:
{bg_context}

Behavioral profile:
{structured_profile}

Instructions:
- Answer as the person.
- Use the profile and background data to infer the most likely
  response.
- If the answer is uncertain, choose the response most consistent
  with the available evidence.
- Provide exactly one answer per question.

Return valid JSON only. No markdown, code fences, or extra text.

Questions:
{questions_text}
```

B.2.4 Retrieval-Augmented Agents: Prediction Prompt

Both retrieval-augmented agents (*Profile + Lexical Top-K* and *Profile + Semantic Top-K*) use an identical prediction prompt; they differ only in how the K retrieved rows in `{related_rows_text}` are selected (Sections B.3–B.4):

```
You are simulating a specific person answering a survey.

Answer each question as this person would, using:
  1. the behavioral profile below
  2. the retrieved prior answers as behavioral evidence
  3. the respondent background data as supporting context

Background data:
{bg_context}

Behavioral profile:
{structured_profile}

Retrieved prior answers (most relevant to this batch):
{related_rows_text}

Instructions:
- Answer as the person.
- Use the profile, background data, and retrieved answers to
  infer the most likely response.
- If the answer is uncertain, choose the response most consistent
  with the available evidence.
```

```
- Provide exactly one answer per question.

Return valid JSON only. No markdown, code fences, or extra text.

Questions:
{questions_text}
```

B.3 Lexical Retrieval

The *Profile + Lexical Top-K* agent ranks prior answered rows by surface-form overlap with the prediction sub-batch.

Tokenization. For each question — whether a target question or a candidate prior row — the `variable_label` and `categories` fields are concatenated, lowercased, and tokenized on non-alphanumeric boundaries. Tokens of length ≤ 1 and tokens belonging to a 40-token stopword list are discarded; the remaining tokens form a set.

Pairwise scoring. For each candidate prior row, we count token overlap with the target question separately over the question label (weighted $\times 3$) and the category field (weighted $\times 1$), then add a bonus of $+2$ if the two questions share the same representation type and $+1$ if they share the same question type.

Batch-level aggregation and selection. Each candidate prior row is scored against the sub-batch by summing its pairwise relevance scores across all target questions in that sub-batch. Rows are ranked by this aggregate score in descending order, with ties broken by original row order, and the top $K = N_{\text{batch}}^{(s)}$ rows are appended to the prediction prompt.

B.4 Semantic Retrieval

The *Profile + Semantic Top-K* agent replaces the lexical scorer with cosine similarity over dense embeddings.

Precomputation. Embeddings for all prior-answer row metadata and all target-question metadata are computed offline using OpenAI `text-embedding-3-small`. Each entry is keyed by `variable_name` and stored in a sibling JSON file alongside the corresponding metadata CSV.

Runtime retrieval. At inference time, each candidate prior row is scored against the sub-batch by summing its cosine similarity to every target question embedding in that sub-batch. Rows are ranked by this aggregate similarity in descending order, and the top $K = N_{\text{batch}}^{(s)}$ are appended to the prediction prompt, following the same selection procedure as the lexical agent.

B.5 Answer Validation and Retry

Every model response undergoes a four-stage validation pipeline before results are written to disk.

1. **Parsing.** Optional Markdown code fences are stripped; the remaining text must parse as valid JSON.
2. **Schema.** The parsed object must contain a `predictions` array whose entries each carry `variable_name` and `predicted_answer` keys.
3. **Coverage.** The set of returned `variable_name` values must exactly equal the expected set for the sub-batch; both omissions and unexpected additions are treated as failures.
4. **Type and range.** Each answer is normalized by declared response type: categorical answers must be a valid category code; numeric answers must fall within the declared range; open-ended answers are accepted as-is.

On failure, the invalid response and the associated error message are appended to the prompt, and the call is reissued. Prediction calls are retried up to four times; profile-generation calls up to three times.

Any sub-batch that exhausts all retries is recorded as a missing prediction and counted as incorrect in accuracy computation.

B.6 Models

Three prediction models are evaluated: **GPT-5.4** (OpenAI), **Gemini-3-Flash-Preview** (Google), and **Claude Haiku 4.5** (Anthropic). All are queried without explicit chain-of-thought reasoning. Profile generation uses **GPT-5.4** exclusively across all agent configurations and both prediction tasks; differences in prediction accuracy across models therefore reflect the prediction model alone, not variation in the persona representation each receives. Semantic retrieval embeddings are produced by the OpenAI `text-embedding-3-small` model.

B.7 Clarification on Baselines

Because our goal is prospective survey approximation, we use baselines that do not require observing human responses to the target questions. We therefore do not include empirical majority-class, demographic-cell majority, or other target-distribution baselines as deployable baselines: these methods require access to the same held-out human answers that the digital personas are intended to approximate. Such comparisons would be oracle references rather than usable baselines in the intended setting. Our no-context LLM baseline instead measures how well the model can reproduce target response distributions without any respondent-specific information, thereby isolating the incremental value of persona conditioning under the same information constraints faced at deployment.

We also distinguish baselines from post-processing strategies applied to persona outputs. Procedures such as majority voting across repeated persona generations, ensembling predictions from multiple persona settings, or calibrating generated responses after prediction may improve downstream estimates, but they operate on the outputs of the persona system rather than providing an independent comparison method. We therefore treat such aggregation or calibration procedures as post-processing choices and leave them outside the scope of the present evaluation framework, which focuses on the reliability of the persona constructions themselves.

C Additional Experiments and Results

C.1 Supplementary Results for Section 5

C.1.1 Detailed Results of All Dimensions

Tables 2 and 5 report the full per-dimension breakdown of persona fidelity for single-wave and core predictions across five persona conditions and three LLMs. Six metrics are included: Clustering, Respondent, Question, Ques. Dist. JSD, Resp. Dist. MMD, and Equity Bias. All values are mean $\pm$ bootstrap SE over $N{=}100$ respondent resamples, with the Baseline condition restricted to GPT. Bold indicates the best value per metric.

- **Clustering (ARI).** K-means is applied independently to the ground-truth response matrix $\mathbf{X}$ and the predicted matrix $\hat{\mathbf{X}}$, yielding cluster assignments vectors $\mathbf{c}$ and $\hat{\mathbf{c}}$. The Adjusted Rand Index (ARI) measures the agreement between the two partitions:

$$\mathrm{ARI}(\mathbf{c},\,\hat{\mathbf{c}}) = \frac{\mathrm{RI} - \mathbb{E}[\mathrm{RI}]}{\max(\mathrm{RI}) - \mathbb{E}[\mathrm{RI}]}, \tag{2}$$

  where RI is the Rand Index. Higher is better (↑).

- **Question Distribution (JSD).** The Jensen–Shannon distance between the predicted response distribution $\hat{P}_q$ and the ground-truth distribution P_q for question q, averaged over all questions:

$$\overline{\mathrm{JSD}} = \frac{1}{Q}\sum_{q=1}^{Q}\sqrt{\frac{1}{2}\mathrm{KL}(P_q\|M_q) + \frac{1}{2}\mathrm{KL}(\hat{P}_q\|M_q)}, \tag{3}$$

  where $M_q = \frac{1}{2}(P_q + \hat{P}_q)$. Lower is better (↓).

- **Respondent Distribution (MMD).** The Maximum Mean Discrepancy between the true respondent embedding matrix $\mathbf{X}$ and the predicted matrix $\hat{\mathbf{X}}$, computed via an RBF kernel k:

$$\mathrm{MMD} = \sqrt{\frac{1}{2}\big(\mathbb{E}[k(\mathbf{x}, \mathbf{x}')] + \mathbb{E}[k(\hat{\mathbf{x}}, \hat{\mathbf{x}}')] - 2\,\mathbb{E}[k(\mathbf{x}, \hat{\mathbf{x}})]\big)}. \tag{4}$$

  Lower is better (↓).

- **Equity (Mean Absolute deviation of DPI).** The mean absolute deviation of the Demographic Parity Index across demographic subgroups $g \in \mathcal{G}$ (gender, age, household stage):

$$\mathrm{EB} = \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} \left| \frac{\mathrm{ACC}_g}{\mathrm{ACC}_{\mathrm{all}}} - 1 \right|, \tag{5}$$

  where ACC_g is respondent-level accuracy for subgroup g and $\mathrm{ACC}_{\mathrm{all}}$ is the overall accuracy. A value of 0 indicates perfect demographic parity. Lower is better (↓).

Compared to the radar plots in Figure 2, which provide an aggregate visual overview of each condition's fidelity profile, the tables supply exact numerical values and bootstrap uncertainty estimates that the radar cannot convey. Both representations converge on the same conclusion: persona-rich conditions (*Profile*, *Profile+Lexical*, and *Profile+Semantic*) consistently outperform *Baseline* across most fidelity dimensions.

We further examine whether the aggregate reliability patterns vary by study domain. Figures 7 and 8 report per-study radar plots for the core-study and single-wave prediction tasks, respectively. These plots show that the main conclusions are not driven by a small number of domains. Across studies, persona conditioning most consistently improves distributional alignment, while exact-match performance changes more modestly and clustering agreement remains weak. The per-study results therefore reinforce the broader aggregated findings.

C.1.2 Additional Details on Figure 4

This heatmap examines where the best-performing digital-persona setting succeeds or fails across two forms of prediction difficulty. The plotted setting is *Profile + semantic top-k* with *Gemini 3 Flash* for Single Wave.

The x-axis captures the respondent's answer pattern. For each question, we first calculate how common each real human answer is. For a respondent-question pair, answer rarity is defined as

$$\text{answer rarity} = 1 - \text{share of respondents who gave the same answer.}$$

We then average this rarity score across all answered questions for each respondent. Respondents are divided into five rank-based bins, from *Common* to *Rare*. Thus, respondents on the left tend to give common human answers, while respondents on the right tend to give rarer answer patterns.

The y-axis captures variation in the real human answers for each question. For each question, we compute the entropy of the human answer distribution and normalize it by the maximum possible entropy for that question:

$$\text{answer variability} = \frac{\text{entropy of human answer distribution}}{\text{maximum possible entropy}}.$$

Questions are then divided into five rank-based bins, from *Low* to *Very high*. Low-variation questions are those where humans mostly agree, while very-high-variation questions are those where humans are highly split across answer options.

Each heatmap cell combines one respondent answer-pattern bin with one question-variation bin. Within each cell, performance is computed using question-level F1 scores and then averaged across questions in that cell. This avoids letting cells with more respondent-question rows dominate only because they contain more observations. The color scale ranges from 0 to 1, with darker colors indicating higher predictive performance.

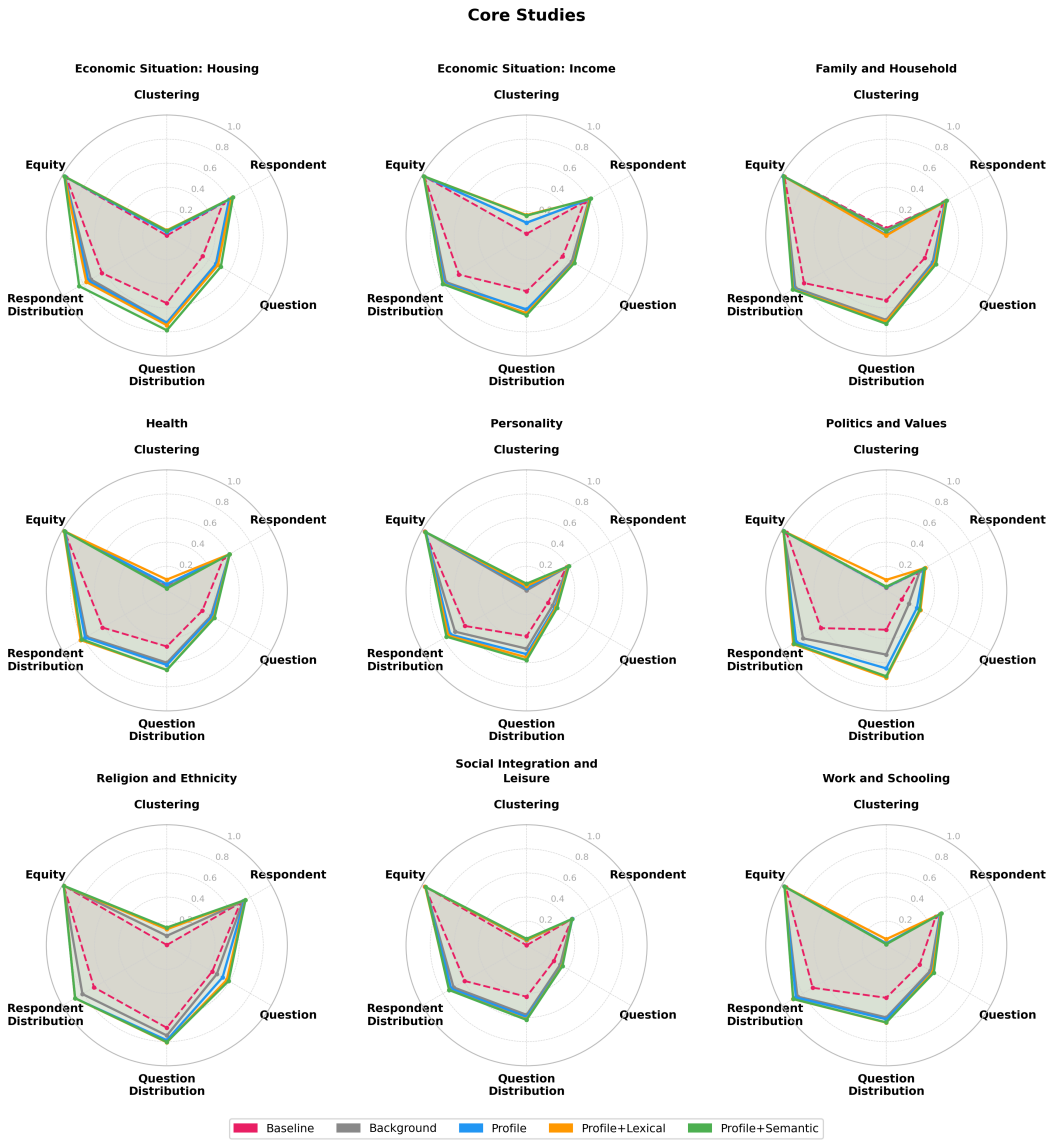


Figure 7: Per-study radar plots for core-study prediction across nine study domains. Each radar summarizes performance across the six evaluation dimensions for all persona settings. Distributional alignment gains from persona conditioning are visible across most domains, but are largest for Religion and Ethnicity and Economic Situation studies. Clustering performance remains near zero across all domains and all settings, confirming that the failure to recover respondent cluster structure is not domain-specific but a consistent limitation of persona-generated responses.

The heatmap shows that performance is strongest in the lower-left region, where questions have low human-answer variation and respondents have common answer patterns. For example, the *Low* variation and *Common* respondent-pattern cell reaches approximately 0.75. Performance declines as questions become more variable and as respondent answer patterns become rarer, indicating that digital personas are most reliable for ordinary questions and ordinary respondents, but less reliable for heterogeneous questions or respondents with uncommon answer profiles.

C.1.3 Clustering

For the clustering analysis, each respondent was treated as a multivariate answer profile rather than evaluating one question at a time. For each analysis unit, respondents' real answers and corresponding predicted answers were encoded into the same feature space. Binary and multiple-choice questions

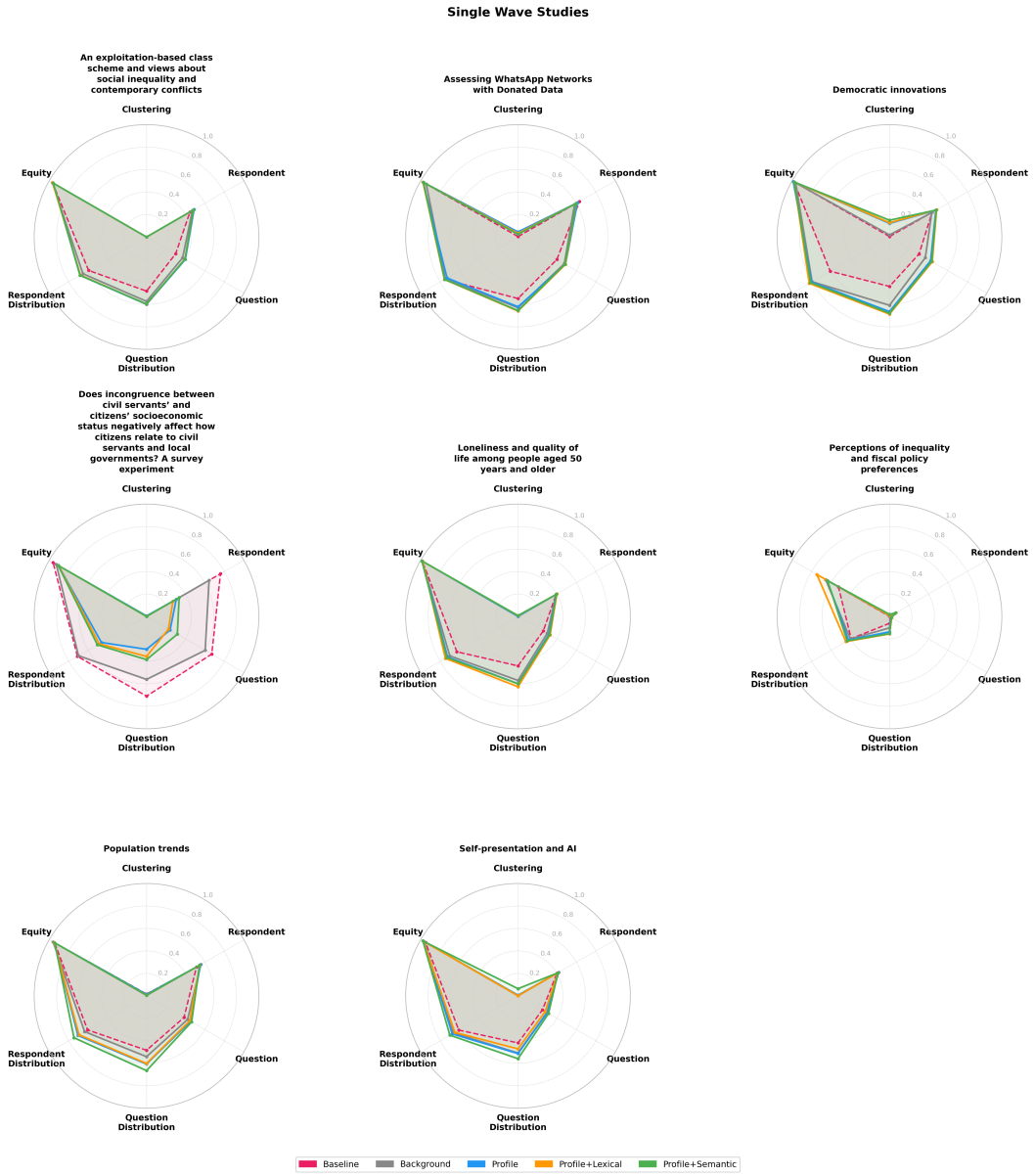


Figure 8: Per-study radar plots for single-wave prediction across eight study domains. The pattern of results mirrors core studies: distributional alignment shows the clearest improvement over the no-context baseline across domains, while exact-match and equity metrics remain broadly stable. Clustering agreement is consistently weak regardless of study domain or persona architecture, reinforcing the finding that persona-generated responses do not reliably recover the respondent grouping structure present in human data. Studies with more concentrated answer distributions, such as Democratic Innovations, show comparatively stronger question-level match performance.

were represented as categorical answer features. We then used k-means clustering to partition respondents into groups based on the similarity of their answer profiles. We used up to $k = 7$ clusters based on the silhouette score.

The same clustering logic is used across all clustering figures, but the analysis unit changes depending on the plot. For each digital-persona architecture and large language model, clustering is performed separately on that setting's predicted answers and compared to the real-human clustering using the adjusted Rand index (ARI). ARI measures whether the predicted-answer clusters group respondents in the same way as the real-human clusters.

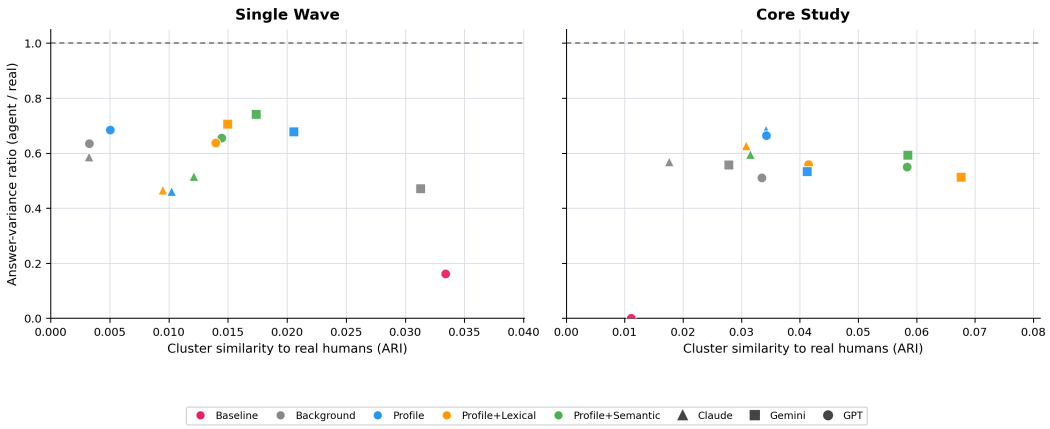


Figure 9: Clustering similarity and answer-variance preservation for agent-generated responses. Each point represents one digital persona configuration, with color indicating the prompting architecture and shape indicating the large language model. The x-axis reports similarity between agent and human cluster structure using ARI; higher values indicate closer recovery of human clustering. The y-axis reports the ratio of answer variance in agent responses relative to real human responses, where 1.0 indicates equal variance and lower values indicate compression of human response diversity. Panels show results separately for Single Wave and Core Study samples.

Figure 9 summarizes this comparison at the model-setting level. The Single Wave panel has lower ARI overall than the Core Study panel, with most points below about 0.035. However, several Single Wave settings have relatively high answer-variance ratios, around 0.6 to 0.75. This suggests that some settings preserve the amount of answer variation better than they preserve the actual grouping of respondents. The Core Study panel shows somewhat higher ARI values, reaching around 0.06 to 0.07 for some settings. But the answer-variance ratios remain below 0.7, meaning that even when the clustering structure is somewhat closer to humans, the predicted answers still compress human variation. The main observation is that digital personas do not preserve clustering structure equally across topics.

For the demographic breakdown, clustering is recomputed separately within each demographic subgroup. Figure 10 shows that clustering preservation varies strongly by subgroup. In the Single Wave panel, gender and household-stage groups have noticeably higher ARI than many age groups. Female and Male subgroups have the highest ARI values, often around 0.4 to 0.6 depending on setting, while the 50-64 age group is much lower. In the Core Study panel, ARI values are generally much lower than in the Single Wave demographic panel. Most age-group values are close to zero, and household-stage values are also low. The clearest exception is gender, especially Male, where several settings reach higher ARI values than other Core Study subgroups. So the takeaway is that predicted answers recover demographic subgroup structure unevenly.

For the study-level plots, clustering is recomputed separately within each study or domain. In Figure 11, digital personas do not preserve clustering structure equally across topics. Clustering recovery is highest in Family and Household, where some settings reach around 0.6 to 0.7 ARI. Work and Schooling has moderate ARI values, roughly around 0.3 to 0.4 for several settings. Economic Situation: Housing is lower, around 0.15 to 0.25. In Figure 12, clustering preservation is highly study-specific. Assessing WhatsApp Networks with Donated Data. This study has much higher ARI than the others, with the best setting reaching around 0.35. Several model/settings in this study also cluster around 0.2 to 0.26, which is still far above the rest of the Single Wave studies. Most other Single Wave studies have ARI values near zero or only slightly above zero. The takeaway is that Single Wave clustering preservation is highly study-specific.

C.2 Supplementary Experiments for Section 6

To comprehensively assess digital persona prediction accuracy, we fitted the full modeling pipeline under both prediction directions: single-wave prediction and core prediction. For each direction, models were estimated across three feature layers: *behavioral* (including per-question answer variability

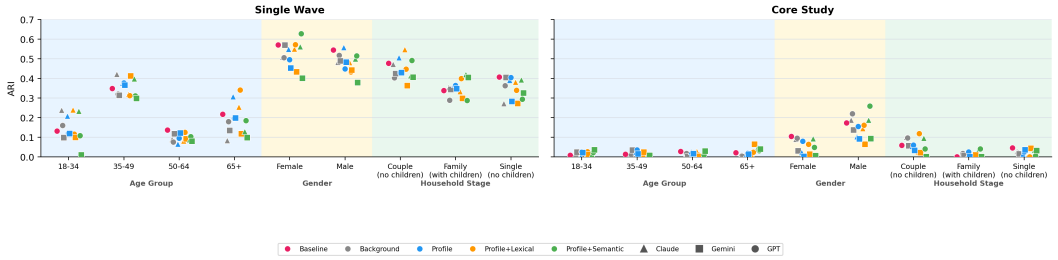


Figure 10: Clustering similarity by demographic subgroup. Each point reports the adjusted Rand index between real-human clusters and agent-generated clusters within a demographic subgroup. Clustering is recomputed separately for each subgroup rather than estimated once on the full sample. Colors indicate agent architecture and shapes indicate language model. Higher ARI values indicate stronger preservation of human multivariate response structure.

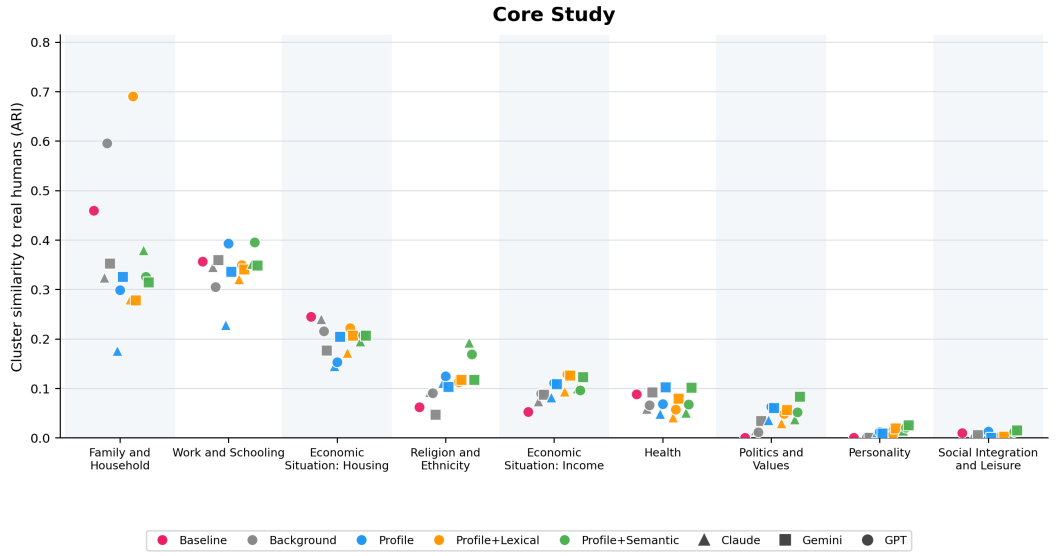


Figure 11: Core Study clustering similarity by topic area. Points show adjusted Rand index between real-human clusters and agent-generated clusters within each Core Study domain. Each domain is clustered separately using the respondent answer profiles available for that domain. Colors indicate agent architecture and shapes indicate language model. Higher values indicate stronger recovery of human response-pattern structure.

and per-respondent answer rarity), *contextual* (retaining study domain but removing behavioral aggregates), and *structural* (question format and demographics only, excluding study domain). Within each layer, five model classes were applied: Logistic Regression with clustered standard errors, Mixed-Effects Logistic Regression (MEGLM) with crossed random intercepts for respondent and question, Decision Tree (max depth = 5), Random Forest ($n = 300$ trees), and XGBoost ($n = 400$ estimators) with SHAP-based feature attribution.

Figure 13 shows SHAP beeswarm plots for the XGBoost model under single-wave prediction across all three feature layers. Answer variability and the number of response options dominate the behavioral layer, while question format and demographic variables emerge as the primary drivers in the structural layer, a pattern consistent with the core-prediction results in Figure 6.

Tables 3 and 4 summarise the full results. Across both prediction directions, the behavioral layer performs strongest overall, with XGBoost achieving the highest accuracy. The gap between behavioral and structural layers is more pronounced in the direction of core prediction, where richer feature sets confer a larger marginal gain.

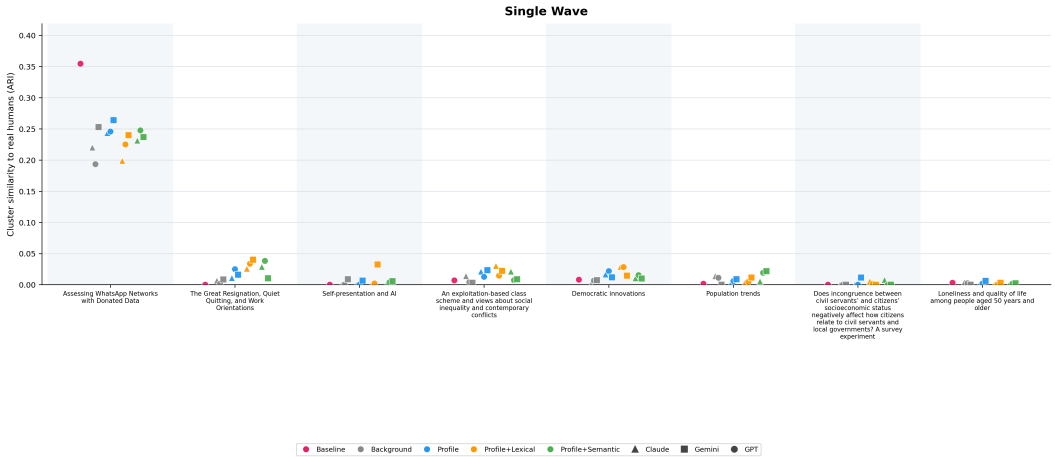


Figure 12: Single Wave clustering similarity by study. Points show adjusted Rand index between real-human clusters and agent-generated clusters for each Single Wave study. Clustering is computed separately within each study using respondent-level answer profiles. Colors indicate agent architecture and shapes indicate language model. Higher ARI values indicate stronger preservation of human multivariate response structure.

Table 4: Single-wave prediction performance across three feature layers ($n = 30{,}725$).

	Behavioral		**Contextual**		**Structural**	
Model	ACC	AUC	ACC	AUC	ACC	AUC
Logistic Regression	0.671 ($\pm$0.003)	0.744	0.618 ($\pm$0.003)	0.692	0.612 ($\pm$0.003)	0.611
MEGLM	0.670 ($\pm$0.003)	0.744	0.617 ($\pm$0.003)	0.692	0.612 ($\pm$0.003)	0.607
Decision Tree	0.685 ($\pm$0.003)	0.757	0.637 ($\pm$0.003)	0.699	0.625 ($\pm$0.003)	0.667
Random Forest	0.715 ($\pm$0.003)	0.773	0.679 ($\pm$0.003)	0.737	0.647 ($\pm$0.003)	0.686
XGBoost	**0.761** ($\pm$0.003)	**0.762**	**0.693** ($\pm$0.003)	**0.725**	**0.652** ($\pm$0.003)	**0.670**

D Sampling Design

We construct separate respondent samples for the two prediction tasks, which we refer to as the Core and Single-Wave samples. In both cases, the goal is to preserve demographic variation while selecting respondents with sufficient prior and target answer coverage. Table 6 is the summary of two sample groups.

Eligible population. For each task, a respondent is eligible if they have a non-empty background file, non-empty prior-answer history for the relevant input scope, and at least one evaluable target answer. Let U denote the eligible respondent population for a task, with size N. Each respondent $i \in U$ is associated with demographic variables and response-count variables, including the number of available prior answers n_i^{prior} and target answers n_i^{target}.

Demographic strata. We stratify respondents by age group, gender, and household stage. Age is grouped into four categories: 18–34, 35–49, 50–64, and 65+. Household stage is derived from partner status, children, and living situation, and grouped into family with children, couple without children, single without children, and other or unknown. The primary stratum is the joint demographic cell

$$h = \text{Age group} \times \text{Gender} \times \text{Household stage}.$$

Sparse strata with fewer than five eligible respondents are excluded from allocation. Table 7 reports the available and allocated counts for each stratum in both samples.

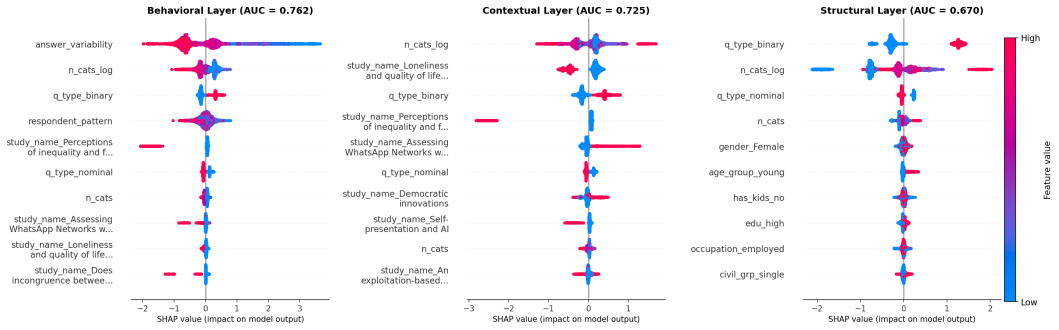


Figure 13: SHAP beeswarm plots showing the top ten predictors of digital persona accuracy (single-wave prediction) across three feature layers.

Table 5: Fidelity of digital personas on single-wave predictions across five persona conditions and three LLMs. Metrics and notation follow Table 2. Bold denotes values statistically tied with the best-performing setting under overlap of the reported uncertainty intervals.

		Baseline	Background	Profile	Profile+Lexical	Profile+Semantic
Clustering ARI (↑)	**GPT**	0.067 (±0.014)	0.044 (±0.017)	0.145 (±0.020)	0.127 (±0.026)	0.169 (±0.026)
	Claude	—	0.076 (±0.029)	**0.316 (±0.056)**	0.216 (±0.037)	0.209 (±0.034)
	Gemini	—	0.064 (±0.024)	0.125 (±0.020)	0.129 (±0.023)	0.126 (±0.021)
Resp. Match rate (↑)	**GPT**	0.445 (±0.003)	0.446 (±0.003)	0.455 (±0.003)	0.451 (±0.003)	0.452 (±0.003)
	Claude	—	0.430 (±0.003)	0.442 (±0.003)	0.442 (±0.003)	0.443 (±0.003)
	Gemini	—	0.458 (±0.003)	**0.467 (±0.003)**	0.459 (±0.003)	**0.462 (±0.003)**
Question F1 score (↑)	**GPT**	0.303 (±0.004)	0.358 (±0.004)	0.373 (±0.003)	0.373 (±0.004)	0.374 (±0.003)
	Claude	—	0.347 (±0.004)	0.359 (±0.003)	0.358 (±0.003)	0.366 (±0.003)
	Gemini	—	0.361 (±0.005)	**0.379 (±0.004)**	**0.378 (±0.004)**	**0.382 (±0.004)**
Ques. Dist. JSD (↓)	**GPT**	0.537 (±0.003)	0.416 (±0.004)	0.407 (±0.004)	**0.394 (±0.004)**	**0.399 (±0.004)**
	Claude	—	0.435 (±0.003)	0.415 (±0.004)	0.420 (±0.004)	**0.399 (±0.004)**
	Gemini	—	0.451 (±0.004)	0.417 (±0.004)	0.408 (±0.004)	0.404 (±0.004)
Resp. Dist. MMD (↓)	**GPT**	0.292 (±0.003)	0.235 (±0.003)	0.230 (±0.003)	**0.216 (±0.003)**	0.225 (±0.003)
	Claude	—	0.254 (±0.003)	0.236 (±0.002)	0.231 (±0.002)	0.230 (±0.002)
	Gemini	—	0.249 (±0.003)	0.235 (±0.003)	0.225 (±0.003)	0.227 (±0.003)
Equity DPI (↓)	**GPT**	**0.021 (±0.004)**	**0.029 (±0.004)**	**0.024 (±0.004)**	**0.023 (±0.004)**	**0.021 (±0.003)**
	Claude	—	0.032 (±0.005)	**0.026 (±0.004)**	**0.022 (±0.004)**	**0.027 (±0.004)**
	Gemini	—	**0.027 (±0.004)**	**0.025 (±0.004)**	**0.027 (±0.005)**	**0.028 (±0.004)**

Proportional allocation. For each task, we select $n = 500$ respondents. Let N_h be the number of eligible respondents in stratum h, with $N = \sum_h N_h$. The target allocation for stratum h is

$$a_h^* = \frac{N_h}{N} n.$$

Because a_h^* is generally non-integer, we use largest-remainder proportional allocation. Each stratum first receives $\lfloor a_h^* \rfloor$ respondents, and remaining slots are assigned to strata with the largest fractional remainders. Allocations are capped by the available stratum size, so the final allocation satisfies

$$\sum_h n_h = n, \qquad 0 \le n_h \le N_h.$$

Within-stratum selection. Within each stratum, respondents are ranked by answer coverage. For respondent i in stratum h, define

$$x_i = n_i^{\text{prior}}, \qquad y_i = n_i^{\text{target}}.$$

We compute within-stratum standardized scores

$$z_i^{(x)} = \frac{x_i - \bar{x}_h}{s_{x,h}}, \qquad z_i^{(y)} = \frac{y_i - \bar{y}_h}{s_{y,h}},$$

where $\bar{x}_h$ and $\bar{y}_h$ are stratum means, and $s_{x,h}$ and $s_{y,h}$ are stratum standard deviations. If a standard deviation is zero or undefined, the corresponding standardized score is set to zero. The main ranking

Table 6: Demographic composition of the two respondent samples.

Category	Group	Single-wave	Core
Gender	Female	266	271
	Male	234	229
Age group	18–34	19.0%	21.2%
	35–49	20.2%	20.4%
	50–64	28.4%	27.4%
	65+	32.4%	31.0%
Household stage	Couple without children	39.4%	38.0%
	Family with children	35.0%	36.4%
	Single without children	25.6%	25.6%

score is

$$s_i = z_i^{(x)} + z_i^{(y)}.$$

Ties are broken using the product $x_i y_i$, followed by the target and prior answer counts. Before ranking, respondents within each stratum are randomly shuffled with a fixed seed to make tie-breaking reproducible. The top n_h respondents in each stratum are selected.

Allocation accuracy. We evaluate how closely each selected sample preserves the demographic composition of the eligible population. For demographic cell c, let Sample_c be the selected count and

$$\mathrm{Expected}_c = \frac{\mathrm{Population}_c}{\sum_j \mathrm{Population}_j} n$$

be the proportional target count. We summarize allocation error using mean absolute discrepancy,

$$\mathrm{MAD} = \frac{1}{C} \sum_{c=1}^{C} \left| \mathrm{Sample}_c - \mathrm{Expected}_c \right|,$$

and maximum absolute discrepancy,

$$\mathrm{MaxD} = \max_{1 \le c \le C} \left| \mathrm{Sample}_c - \mathrm{Expected}_c \right|.$$

As shown in Table 8, both samples have $\mathrm{MAD} = 0.2602$ and $\mathrm{MaxD} = 0.5275$. Thus, the selected samples closely preserve the eligible population distribution over age group, gender, and household stage, with an average deviation of less than one respondent per demographic cell.

Final distribution of respondents. Figure 14 summarizes respondent-level answer coverage in the prior and target partitions used for evaluation after sampling.

E Positioning Relative to Prior Work (Expanded)

LLMs as synthetic survey respondents. A growing body of work asks whether LLMs can approximate human survey responses and support scalable social measurement. Early studies suggested that LLMs conditioned on demographic profiles can reproduce some population-level response patterns [Argyle et al., 2023]. Subsequent work has offered a more cautious picture: synthetic responses may distort variation and statistical relationships even when aggregate averages appear plausible [Bisbee et al., 2024], and LLM survey answers can be sensitive to ordering, labeling, and other design artifacts [Dominguez-Olmedo et al., 2024]. More recent work has explored whether model adaptation can improve aggregate survey simulation, including specialized models for global response distributions and personalized survey-data modeling [Cao et al., 2025, Kaiser et al., 2025]. Together, these studies show that LLMs can sometimes approximate population-level patterns, but aggregate alignment alone does not establish whether simulated respondents preserve individual answers, subgroup relationships, or scientific conclusions.

Table 7: Allocation counts by demographic stratum for the two prediction tasks.

Age group	Gender	Household stage	Single-Wave		Core	
			Available	Allocated	Available	Allocated
18–34	Female	Couple without children	90	12	159	14
		Family with children	220	28	338	29
		Single without children	91	12	146	13
	Male	Couple without children	60	8	98	9
		Family with children	159	21	245	21
		Single without children	68	9	104	9
	Other	Other or unknown	–	–	5	0
35–49	Female	Couple without children	72	9	88	8
		Family with children	305	39	419	37
		Single without children	77	10	128	11
	Male	Couple without children	63	8	104	9
		Family with children	230	30	298	26
		Single without children	82	11	113	10
50–64	Female	Couple without children	277	36	362	32
		Family with children	248	32	327	28
		Single without children	126	16	169	15
	Male	Couple without children	231	30	294	26
		Family with children	224	29	297	26
		Single without children	136	18	174	15
	Other	Other or unknown	–	–	1	0
65+	Female	Couple without children	288	37	495	43
		Family with children	18	2	36	3
		Single without children	227	29	375	33
	Male	Couple without children	406	52	642	56
		Family with children	33	4	56	5
		Single without children	136	18	254	22

Table 8: Allocation accuracy for the two respondent samples.

Prediction task	Eligible population	Sample size	MAD	MaxD
Core	4,266	500	0.2602	0.5275
Single-Wave	5,785	500	0.2602	0.5275

Grounded digital personas and personalized user modeling. A related literature moves beyond generic demographic prompting toward personas grounded in richer individual information. Prior opinions can improve prediction beyond demographics and ideology alone [Hwang et al., 2023], and recent digital-persona systems condition models on combinations of demographic, behavioral, psychometric, conversational, or self-report data [Chen et al., 2025, Jiang et al., 2025, Park et al., 2026, Toubia et al., 2025]. These studies show that richer personal histories can improve persona fidelity and make individual-level simulation more plausible. Related meta in-context learning methods learn across groups to align an LLM with group-level preferences, whereas our setting is training-free and evaluates whether personas grounded in individual respondents' histories preserve held-out survey findings [Zhao et al., 2024]. However, prior persona studies leave open which forms of personal history, profile construction, and retrieved memory are useful for predicting held-out responses from the same real individuals, especially in fixed-choice longitudinal survey settings.

Evaluating persona fidelity and scientific reliability. Evaluations of digital personas vary widely, including aggregate distribution matching, personalized response selection, behavior-chain prediction, consistency evaluation, persona-effect measurement, validation of LLM simulations, and human-likeness judgments [Jiang et al., 2025, Li et al., 2025c,b, Abdulhai et al., 2026, Hu and Collier, 2024, Hullman et al., 2026]. These benchmarks are valuable, but scientific use of digital personas requires

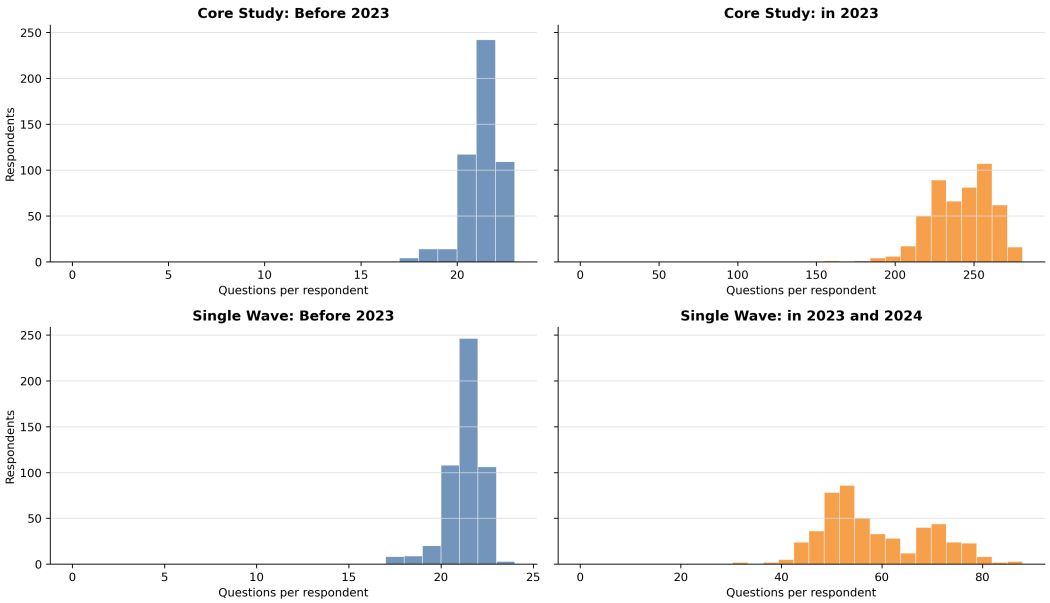


Figure 14: Distribution of the number of available questions per respondent across the four survey partitions used in the prediction tasks. The left panels show pre-2023 prior-answer coverage for core-study and single-wave histories, while the right panels show held-out target-answer coverage for core-study targets in 2023 and single-wave targets in 2023–2024. The histograms indicate that sampled respondents have dense prior-answer histories and substantial held-out target coverage, supporting respondent-level evaluation.

a stricter test: predictions should be evaluated against real held-out responses from the people the personas are meant to represent. A persona may generate plausible or personalized outputs while still failing to preserve individual answers, question-level reliability, subgroup patterns, or study-level conclusions. This motivates evaluation across target question types, information sources, persona architectures, model families, and survey domains.

Limits, biases, and conditional substitution. Recent work also cautions against treating LLM personas as direct substitutes for human participants. LLMs may misportray or flatten identity groups [Wang et al., 2025], introduce systematic downstream biases through persona generation [Li et al., 2025a], or produce heterogeneous effects across items and subgroups [Morocho et al., 2026]. These findings suggest that digital personas should be treated as conditional scientific instruments rather than universal replacements for human data.

Our study follows this view by asking when digital personas grounded in real respondents' prior survey histories can predict held-out responses from the same individuals, and when their failures make real human data indispensable. Table 1 positions our study against prior work along three dimensions: the data setting, the persona construction strategy, and the evaluation design. We distinguish studies that use objectively scored fixed-choice outcomes, representative stratified samples, same-person temporal holdouts, real longitudinal panels, and prior survey histories; studies that compare persona variants, use retrieval memory, evaluate multiple LLMs, and include a no-context baseline; and studies that assess individual fidelity, aggregate distributions, subgroup equity, joint response structure, and error drivers.

The main pattern is that prior work covers important but partial areas of this design space. Appendix E.1 defines the comparison criteria, and Appendix E provides the full related-work discussion and study-by-study comparison. Several of the closest studies are recent preprints that have not yet appeared in peer-reviewed venues, but the overall pattern is consistent: no single study combines the full set of data, persona-construction, and evaluation elements that we examine here.

E.1 Comparison criteria

The comparison in Table 1 uses fifteen criteria that capture the main methodological and evaluative components of our study.

Fixed-choice answers. The target outcomes are constrained to predefined survey response options, such as binary, ordinal, nominal, or multiple-choice categories. This makes evaluation objective and directly comparable across humans and simulated personas, rather than relying only on semantic similarity or subjective plausibility judgments.

Representative stratified population sample. The study uses a sample designed to represent a broader population, with explicit stratification or population-based sampling. This differs from convenience samples, crowdsourced samples, synthetic-only populations, or small purposive samples.

Same-person holdout. The study predicts held-out responses for the same individuals whose prior information is used to construct or condition the persona. This is stricter than predicting aggregate distributions or simulating generic members of demographic groups.

Temporal split. The study separates prior information and target outcomes by time, so that personas are conditioned only on information available before a cutoff and evaluated on later responses. This reduces leakage and better matches real prospective prediction.

Real longitudinal panel. The data come from real respondents observed repeatedly across waves, studies, or time periods. This allows the study to evaluate whether personas preserve individual continuity, rather than only one-time demographic or attitudinal profiles.

Prior survey history. The persona is grounded in respondents' previous survey answers, not only in demographics, short descriptions, interviews, or manually written profiles. This criterion captures whether the model receives structured evidence about the person's past choices.

Persona variants. The study compares more than one way of constructing or prompting personas, such as background-only prompts, generated profiles, profile-plus-memory variants, or other alternative conditioning schemes.

Retrieval memory. The persona has access to a retrieved subset of relevant prior information at prediction time. This differs from giving a static profile or all background information without question-specific retrieval.

Multi-LLM comparison. The study evaluates more than one language model family or model variant, allowing results to distinguish persona-method effects from model-specific effects.

No-context baseline. The study includes a baseline in which the model answers without respondent-specific information. This tests whether persona grounding improves prediction beyond what can be inferred from the question wording and answer options alone.

Individual fidelity. The study evaluates whether simulated personas reproduce the held-out answers of specific individuals. This includes metrics such as exact accuracy, F1, or respondent-level predictive performance.

Aggregate distribution. The study evaluates whether simulated responses reproduce population-level or question-level response distributions, even when individual-level prediction may be imperfect.

Equity / subgroups. The study examines whether performance differs across demographic or social groups, or whether persona simulation introduces, amplifies, or reduces subgroup disparities.

Joint structure. The study evaluates whether simulated responses preserve multivariate relationships across questions or respondents, such as clustering, correlation structure, latent groups, or joint response patterns.

Error drivers. The study analyzes which factors explain prediction errors, such as question type, answer rarity, response entropy, domain, respondent characteristics, or model architecture.

Overall, Table 1 indicates that our contribution is not simply the use of LLMs for survey simulation, nor simply the construction of respondent-level personas. The contribution is the combination of a representative longitudinal panel, prior survey histories, fixed-choice held-out targets, same-person temporal prediction, multiple persona architectures, retrieval-augmented memory, multiple model families, a no-context baseline, and evaluation across individual, aggregate, subgroup, joint-structure, and error-driver dimensions.

F Computational Resources and Cost

Running the full experimental pipeline cost approximately $3,000 in API usage, including $1,700 for GPT-5.4, $600 for Claude Haiku 4.5, and $700 for Gemini 3 Flash Preview. In addition, we used GPT-5.4 to generate respondent profiles at an average cost of approximately $0.30 per respondent. The full run took approximately 80 hours on a single 8-core laptop. For each LLM, we evaluated four persona architectures on 500 respondents in each of the two prediction tasks, yielding 4,000 respondent-architecture-task predictions per LLM. This corresponds to an average API cost of approximately $0.4 per GPT-5.4 prediction, $0.15 per Claude prediction, and $0.18 per Gemini prediction. Equivalently, running all four persona architectures across both tasks costs approximately $3.40 per respondent for GPT-5.4, $1.20 for Claude Haiku 4.5, and $1.40 for Gemini 3 Flash. By construction, token usage increases from Background to Profile and retrieval-augmented architectures as additional respondent context is included. Retrieval-augmented variants provide only modest performance gains over Profile. For larger populations, retrieval costs scale approximately linearly with the number of respondents, while embeddings are computed once.

G Broader Impacts

This work examines when large language models can reliably substitute for human survey respondents, a question with direct implications for social science research, public opinion measurement, and policy-relevant data collection. By identifying the conditions under which digital personas succeed and fail, our findings help researchers make more informed decisions about when synthetic respondents can appropriately supplement or replace costly human data collection. The clearest benefit is distributional approximation for low-variability questions which suggests targeted applications in instrument pre-testing and early-stage population simulation, where approximate aggregate patterns are sufficient and the cost of fielding a full human sample is prohibitive.

At the same time, this work surfaces risks that deserve attention. Digital personas perform poorly for rare, heterogeneous, and idiosyncratic responses which are precisely the kinds of answers that matter most for studying underrepresented groups, minority opinions, and social heterogeneity. Deploying digital personas without accounting for these failure modes risks systematically flattening the very variation that survey research is designed to capture, potentially misrepresenting minority populations or producing misleading subgroup comparisons. Researchers and practitioners should treat digital personas as conditional instruments requiring task-specific validation rather than as general-purpose replacements for human participants, and should exercise particular caution in applications where individual-level fidelity or subgroup accuracy is scientifically or ethically consequential.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: As stated in the question.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes].

   Justification: We include the limitations in Section 7

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A].

Justification: No theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes].

   Justification: All experimental details and results are provided in Section 3, 5–6 and Appendix B–C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes].

Justification: Details and code are available in Appendix B.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes].

   Justification: See experiment setting and details in Section 3, 5–6 and Appendix B–C..

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer:[Yes].

   Justification: We provide specific stratified sampling design with details in Appendix D. We also provide bootstrap confidence intervals for the main results and the modeling in Tables 2, 3, and 4 (in Appendix).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes].

   Justification: We provide a specific section in Appendix F.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes].

   Justification: As stated in the question.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes].

    Justification: This is a subsection in the Appendix G.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A].

    Justification: No risk.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes].

    Justification: We provide details on data usage and license in Appendix A.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer:[N/A].

    Justification: No new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A].

    Justification: No crowdsourcing nor research with human subjects besides usage of already available data.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A].

    Justification: No crowdsourcing nor research with human subjects besides usage of already available data.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes].

Justification: Digital personas are fully LLM dependent, this is stated clearly in the manuscript and further detailed in the Appendix B.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.