SYNTHIA: Scalable Grounded Persona Generation from Social Media Data

Vahid Rahimzadeh*[1,2], Erfan Moosavi Monazzah*[1], Mohammad Taher Pilehvar[3], and Yadollah Yaghoobzadeh[1,2]
[1]Tehran Institute for Advanced Studies, Khatam University, Iran
[2]University of Tehran, Iran
[3]Cardiff University, United Kingdom
{v.rahimzade, e.moosavi_monazzah}@teias.institute

Abstract

Persona-driven simulations are increasingly used in computational social science, yet their validity critically depends on the fidelity of the underlying personas. Constructing virtual populations that are both authentic and scalable remains a central challenge. We introduce SYNTHIA, a persona-generation framework that grounds LLM-generated personas in real social-media posts while delegating narrative construction to language models, using publicly available data from the Bluesky platform. Across multiple social-survey benchmarks, SYNTHIA improves alignment with human opinion distributions over prior state-of-the-art approaches while relying on substantially smaller models. A multi-dimensional fairness and bias analysis shows that SYNTHIA outperforms previous methods for most demographics across different dimensions. Uniquely, SYNTHIA preserves interaction-graph structure among personas grounded in real social network users, enabling network-aware analysis, which we demonstrate through two homophily-focused case studies. Together, these results position SYNTHIA as a practical and reliable framework for constructing scalable, high-fidelity, and equitable virtual populations.

1 Introduction

Persona-driven large language models (LLMs) are increasingly adopted across a wide range of domains (Tseng et al., 2024), particularly in population-level simulation and analysis (Chen et al., 2024b; Xu et al., 2024). In this context, personas may range from simple demographic descriptors to rich psychological profiles and detailed life narratives (Li et al., 2025; Cintas et al., 2025). While explicitly conditioning models on demographic attributes can inadvertently promote stereotypical inferences and amplify bias (Anthis et al., 2025), employing context-rich personas has been shown to foster more individualized variation and reduce disparities in predictive accuracy across demographic groups (Park et al., 2024).

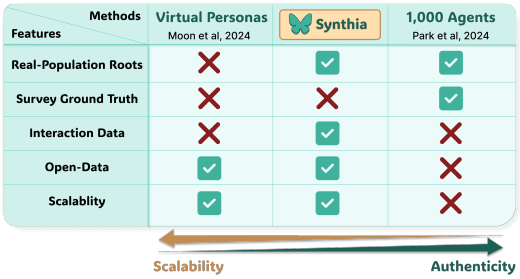


Figure 1: SYNTHIA vs. leading persona methods.

However, scaling the construction of context-rich personas remains a central challenge, with methods falling along a spectrum between authenticity and scalability (see Figure 1). At one extreme are interview-based approaches that derive personas from human data and often yield improved realism (Anthis et al., 2025). Yet, these approaches are resource-intensive and difficult to scale (Park et al., 2024). In contrast, fully synthetic approaches (Moon et al., 2024) offer scalability but frequently introduce systematic artifacts that reduce realism (Li et al., 2025).

Seeking an optimal balance between scalability and authenticity, we introduce SYNTHIA, **S**ynthetic **Y**et **N**aturally **T**ailored **H**uman-**I**nspired Person**A**, a methodology that grounds persona generation in real social media content. Social media contains large volumes of user-generated content reflecting diverse behaviors and viewpoints. For this work, we utilize Bluesky[1], due to its open platform structure and permissive redistribution policies. Concurrently, LLMs have demonstrated a remarkable proficiency in processing such content for persona development (Yin et al., 2025; Prottasha et al., 2025).

SYNTHIA differs from prior work in how it uses language models. A model is used to compose a

*Equal contribution, ordered randomly

[1]https://bsky.app/

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 45983–46006
July 2-7, 2026 ©2026 Association for Computational Linguistics

population of persona narratives from real posts, while separate models conditioned on those narratives answer demographic and opinion questions. We then match the synthetic population to the demographic composition of real survey respondents and evaluate alignment through statistical comparisons between simulated and real-world opinion distributions from social surveys. Thus, the main evaluation is anchored in human survey data rather than LLM-based judgment. An LLM judge is used only for the narrative consistency analysis, where it is validated against human annotations. This design preserves scalability while improving internal factual consistency and reducing systematic bias relative to fully synthetic pipelines, which are key factors for better alignment and fidelity to human populations. Unlike prior works that provide persona text only, SYNTHIA also supplies interaction-graph metadata, enabling network-aware analyses.

Our comprehensive evaluation across 54 experimental configurations establishes SYNTHIA as a robust alternative to SOTA methods. In terms of population opinion alignment, it improves upon baselines by up to 11.6% across social surveys with models less than half the size. Error analysis suggests that these gains are largely driven by a reduction in inner narrative factual contradictions per persona. Furthermore, our fairness analysis reveals that SYNTHIA consistently achieves high fidelity while maintaining stability across demographic groups, reducing accuracy gaps between best and worst performing subgroups by up to 25%, and minimizing disparate impacts in sensitive categories like gender and education. We further demonstrate SYNTHIA's applicability to social network analysis. By preserving the source topology, our personas effectively encode the correlation between structurally and semantically informed homophily, achieving accuracy gains of 8.3% ($p < 0.001$) in link prediction tasks and increasing embedding-space separability between connected and unconnected personas by up to 46%.

Our contributions are fourfold: (i) We propose SYNTHIA, a scalable persona-generation pipeline that produces representative, human-like virtual populations grounded in real social-media content. (ii) We show that internal factual consistency is critical for accurately modeling population-level opinions and diversity. (iii) We demonstrate that grounding reduces systematic bias and improves fairness, enabling reliable persona generation with substantially smaller language models. (iv) We release a large-scale dataset of grounded virtual personas together with their underlying social interaction graph, and illustrate its utility through downstream computational social science case studies.

2 Related Work

Persona-driven use cases of LLMs have been the focus of numerous recent studies (Anthis et al., 2025; Chen et al., 2024b; Tseng et al., 2024; Xu et al., 2024), covering aspects such as the strengths and biases of LLMs (Suh et al., 2025a; Liu et al., 2024; Chen et al., 2024a; Salewski et al., 2023; Cheng et al., 2023; Santurkar et al., 2023; Argyle et al., 2023), computational social science simulations (Piao et al., 2025; Shen et al., 2025; Wang et al., 2025b; Touzel et al., 2024; Rahimzadeh et al., 2025; Argyle et al., 2023), policy and governance decision-making (Piatti et al., 2024; Barnett et al., 2024), and user behavior modeling (Suh et al., 2025b; He et al., 2025; Wang et al., 2025a; He et al., 2024; Park et al., 2023). As the applications of persona-driven models expand, more research has emerged on methodologies for creating these personas (Sun et al., 2025; Bui et al., 2025; Bück-Kaeffer et al., 2025; Liu et al., 2025; Jung et al., 2025; Yin et al., 2025; Dash et al., 2025). Despite current attempts at creating personas through role playing (Chen et al., 2024b; Tseng et al., 2024; Xu et al., 2024), in-context learning (Choi and Li, 2024; Salewski et al., 2023) or aligning models to specific sets of opinions from real users (Hwang et al., 2023; Santurkar et al., 2023), the existing approaches still leave an important gap in scalable methods that combine rich narrative structure with grounding in real user-generated evidence.

One approach is to condition LLMs on backstories that encode a life narrative (Park et al., 2024; Moon et al., 2024). Life narratives provide a structured representation of identity, reflecting demographic and social attributes such as gender, ethnicity, and social class (Moon et al., 2024; Westberg et al., 2024; Stephens and Breheny, 2013). Recent work has emphasized grounding these narratives in real human data to improve authenticity, e.g., Park et al. (2024) simulate the attitudes and behaviors of real individuals by applying LLMs to qualitative interview data and evaluating how well the resulting agents reproduce observed human responses.

Moon et al. (2024) use high-temperature LLM sampling to generate diverse life narratives, exploiting models' broad distributional knowledge.

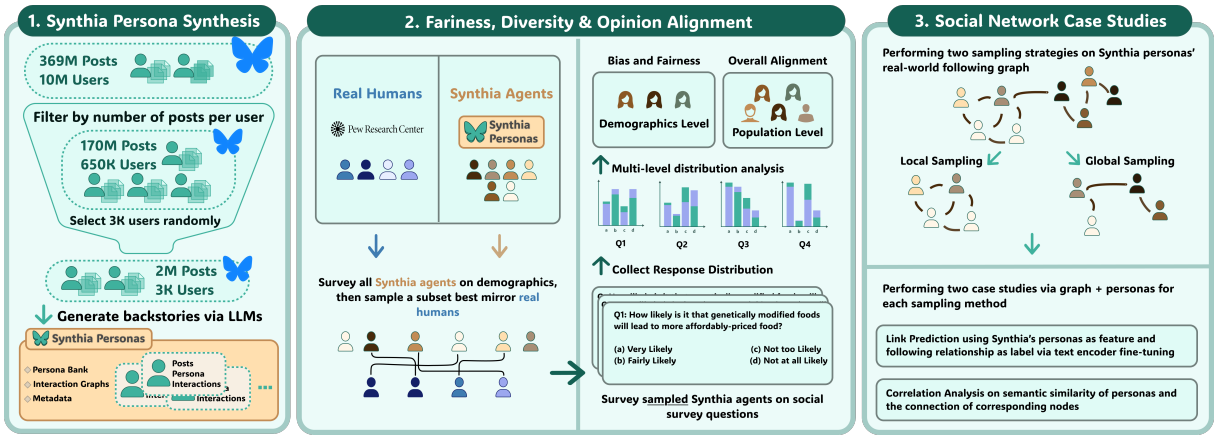


Figure 2: Our approach involves: 1) Collecting and filtering high-quality user data from an open social network and generating SYNTHIA Personas (Sec 3), 2) Evaluating population diversity & opinion alignment with real-world social surveys (Sec 4) and Bias Analysis on the performance and stability across demographics (Sec 5) and 3) Case studies on homophily with social networks (Sec 6).

However, unconstrained generation without real-world grounding risks hallucination and narrative inconsistency.

3 Synthia Persona Synthesis

In this work, we define a persona as a concise first-person life narrative that articulates the backstory of a virtual individual, including preferences, salient life events, and other relevant biographical details. This definition is consistent with prior work on both synthetic (Moon et al., 2024) and human-authored personas (Zhang et al., 2018), enabling direct comparison with existing approaches. Figure 2 gives an overview of our methodology.

User Pool Creation. To ground SYNTHIA personas in real-world data, we curate a diverse pool of social-media content consisting of posts and users from the open platform Bluesky, selected due to its permissive licensing terms that allow public redistribution (see Appendix B). All selected data are normalized into a unified schema, including deduplication and the removal of non-English content. We further filter users with atypical posting behavior by excluding accounts whose activity is either too sparse or too dense. Through preliminary analysis, we observe that users with fewer than 100 posts over a two-year period provide insufficient context, often resulting in overly brief personas or forcing the LLM to hallucinate content. Conversely, users with more than 1,000 posts exceed the context window of the persona generator model (Appendix B). After filtering, our dataset contains approximately 170M posts from 650K unique users. To ensure fair comparison with prior work (Moon et al., 2024; Zhang et al., 2018), we take a random sample of users from this pool with about the same size as the smallest baseline, which is 3K.

Persona Generation. To demonstrate the scalability of our approach, we employ a lightweight open LLM that can be deployed on consumer-grade GPUs. Specifically, we use Gemma-3-27B (Team et al., 2025a). To further assess scalability and analyze the effect of model parameter count on persona backstory generation, we also include a smaller language model (SLM), Phi-4-mini (Team et al., 2025b) (see Appendix A for full experimental details). Before persona generation, we remove explicit social identifiers and interaction cues from the source text, including handles, mentions, URLs, emails, phone numbers. We also exclude replies/reposts from the generation corpus used for persona construction. Using the collected social-media posts for each user, we prompt these models (Figure 9) to generate comprehensive first-person backstories. An example illustrating the grounding between source posts and the resulting persona is shown in Figure 8. To isolate the effect of model size in comparative evaluations, we re-run the persona generation pipeline from the current state-of-the-art synthetic persona framework (Moon et al., 2024) using Gemma-3-27B. While the original backstories in that work were generated using Llama-3-70B, employing the same 27B model across both pipelines ensures a controlled and fair comparison. Dataset statistics for all persona collections are reported in Table 8.

Social Network Graph. A unique characteristic of SYNTHIA is the underlying interaction graph among generated personas. This structure is directly inherited from the original social network of the users selected for synthesis. We formally represent the network as a directed graph, where a directed edge denotes a following relationship between users. This representation enables analyses that jointly consider persona and network structure.

4 Diversity & Opinion Alignment

A primary use of virtual personas is their ability to respond to social surveys as proxies for human respondents. This capability enables the evaluation of fidelity, that is, how faithfully a synthetic population reflects human distributions of social attitudes and behaviors. To assess this alignment, we evaluate personas using surveys from the American Trends Panel (ATP). For direct comparability with prior work, we focus on Wave 34 and Wave 99 of the ATP datasets (Pew Research Center, 2018, 2021). After matching personas to the demographic composition of survey respondents, we compare the simulated opinion distributions of matched personas to real-world survey distributions using statistical metrics. Throughout this section, we use non-instruction-tuned models, as prior work has shown them to outperform instruction-tuned variants in survey-response simulation (Moon et al., 2024). Details regarding survey questions and experimental configurations are provided in Appendix F.3.

4.1 Demographic Matching

To accurately simulate social surveys, the persona population must reflect the demographic composition of the survey respondents. We adopt the demographic matching procedure of Moon et al. (2024), selecting a subset of personas whose aggregate demographic distribution aligns with that of the target population.

For each persona, we infer demographic attributes through *demographic surveying*, in which an LLM is conditioned on the persona's backstory and repeatedly queried with demographic questions. This yields a stable probability distribution over demographic attributes per persona. We then use a greedy matching algorithm to assign each survey respondent to the persona whose inferred demographic profile most closely matches their own. Details of the matching procedure are provided in Appendix F.

Wave	Exp.	EMD ↓	Frob. ↓	Cron. α ↑
W34	$\text{ANT}_{\text{LLaMa}}$	**0.34**	<u>2.41</u>	<u>0.35</u>
	$\text{ANT}_{\text{Gemma}}$	<u>0.34</u>	2.65	0.32
	$\text{SYN}_{\text{Gemma}}$	0.36	**2.25**	**0.38**
	SYN_{Phi}	0.38	2.61	0.31
	$\text{PCHAT}_{\text{Human}}$	0.35	2.76	0.29
	Human	*0.06*	*0.42*	*0.78*
W99	$\text{ANT}_{\text{LLaMa}}$	<u>0.37</u>	**2.03**	**0.41**
	$\text{ANT}_{\text{Gemma}}$	0.49	2.41	0.20
	$\text{SYN}_{\text{Gemma}}$	**0.34**	2.21	0.34
	SYN_{Phi}	0.45	2.39	0.17
	$\text{PCHAT}_{\text{Human}}$	0.58	<u>2.05</u>	<u>0.38</u>
	Human	*0.08*	*0.33*	*0.83*

Table 1: Screening stage results per wave. Best values per wave are **bolded**, second-best are <u>underlined</u>.

4.2 Opinion Alignment Evaluation

After demographic matching, we perform opinion surveying on the matched synthetic population. For each persona, we prompt the LLM to generate responses to the wave-specific opinion questions (see Appendix F.3). To quantify alignment between synthetic and real populations, we compare the resulting opinion distributions using the standard metrics introduced by Moon et al. (2024): Earth Mover's Distance (EMD), Frobenius Norm (Frob.), and Cronbach's Alpha (Cron.). Thus, the main alignment results are obtained by statistical comparison to empirical human response distributions, not by LLM-based judges.

4.3 Experiments and Results

We conduct demographic matching and opinion surveying for five persona sets: (1) $\text{SYN}_{\text{Gemma}}$, our primary SYNTHIA personas; (2) SYN_{Phi}, SYNTHIA personas generated using a smaller model; (3) $\text{PCHAT}_{\text{Human}}$, human-authored personas (Zhang et al., 2018); (4) $\text{ANT}_{\text{Gemma}}$, Anthology personas (Moon et al., 2024) generated using the same model as $\text{SYN}_{\text{Gemma}}$; and (5) $\text{ANT}_{\text{LLaMa}}$, the original Anthology personas.

Screening Stage. Given the high computational cost of full-scale evaluations, we conduct this preliminary screening to filter persona generation methods before proceeding to detailed analysis. To isolate the quality of the personas, we utilize a fixed LLM for response generation across all conditions (see Appendix G). Table 1 presents the results.

We first compare our method against our primary competitor of comparable size, $\textbf{ANT}_{\text{Gemma}}$. $\textbf{SYN}_{\text{Gemma}}$ demonstrates decisive superiority, con-

sistently outperforming $\text{ANT}_{\text{Gemma}}$ in *Frob.* and *Cron.*, which indicates stronger structural alignment and internal consistency. In terms of *EMD*, while $\text{SYN}_{\text{Gemma}}$ maintains parity in Wave 34, it significantly surpasses the competitor in Wave 99 (0.34 vs. 0.49). Remarkably, even our smallest model, $\textbf{SYN}_{\text{Phi}}$, proves comparable to the $\text{ANT}_{\text{Gemma}}$ baseline across waves, despite being generated by a model approximately six times smaller (4B vs. 27B).

Next, we examine the human-generated $\textbf{PCHAT}_{\text{Human}}$ baseline. While $\text{PCHAT}_{\text{Human}}$ performs well in Wave 99, it proves highly volatile compared to the stability of our method. $\text{PCHAT}_{\text{Human}}$ exhibits drastic cross-wave fluctuations (e.g., $\Delta_{\text{Frob.}} = 0.71$) compared to the negligible variance of $\text{SYN}_{\text{Gemma}}$ ($\Delta_{\text{Frob.}} = 0.04$), suggesting that the human-generated baseline suffers from representational inconsistencies.

Finally, $\textbf{SYN}_{\text{Gemma}}$ performs neck-and-neck with the significantly larger state-of-the-art model, $\textbf{ANT}_{\text{LLaMa}}$. Despite $\text{ANT}_{\text{LLaMa}}$ utilizing a generator over twice the size (70B vs. 27B), the leadership is perfectly split: out of the six best performance scores recorded across waves and metrics (3 metrics $\times$ 2 waves), three are secured by $\text{SYN}_{\text{Gemma}}$ and three by $\text{ANT}_{\text{LLaMa}}$. Consequently, we retain $\textbf{ANT}_{\text{LLaMa}}$ and $\textbf{SYN}_{\text{Gemma}}$ as primary candidates, alongside $\textbf{SYN}_{\text{Phi}}$ to analyze the impact of generator scale, allowing us to rigorously evaluate our proposed approach against SOTA methods.

Detailed Analysis. To robustly evaluate the quality of persona sets, we selected three distinct LLMs to serve as both demographic surveyors and response surveyors (setup details in Appendix A). This factorial design, spanning three models for both survey roles, across two waves and three persona sets, yielded a total of 54 experiments ($3 \times 3 \times 2 \times 3$), or 27 per wave. We computed the mean performance per persona set per wave, with a comprehensive overview provided in Table 2.

Crucially, the superior performance of SYNTHIA is robust to the choice of surveyor. When disaggregating results across the factorial design, $\text{SYN}_{\text{Gemma}}$ achieves the top rank in every tested configuration, recording the lowest EMD scores regardless of which model is employed as the Demographic or Response surveyor. For instance, while $\text{ANT}_{\text{LLaMa}}$ and SYN_{Phi} frequently exhibit EMD scores hovering around 0.40, $\text{SYN}_{\text{Gemma}}$ maintains consistently tighter alignment, reaching a minimum EMD of 0.33. Detailed heatmaps visualizing these surveyor-specific dynamics for EMD, Frob, and Cron and the full results across experimental settings are provided in Appendix G.

	Exp.	EMD ↓	Frob. ↓	Cron. α ↑
W34	$\text{ANT}_{\text{LLaMa}}$	0.35 ± 0.02	2.46 ± 0.06	0.34 ± 0.05
	$\text{SYN}_{\text{Gemma}}$	**0.35 ± 0.03***	**2.30 ± 0.20**	**0.39 ± 0.09**
	SYN_{Phi}	0.38 ± 0.04	2.43 ± 0.10	0.38 ± 0.06
	Human	*0.06*	*0.42*	*0.78*
W99	$\text{ANT}_{\text{LLaMa}}$	0.43 ± 0.06	2.14 ± 0.20	0.35 ± 0.10
	$\text{SYN}_{\text{Gemma}}$	**0.38 ± 0.06**	**2.12 ± 0.15**	**0.39 ± 0.09**
	SYN_{Phi}	0.41 ± 0.04	2.21 ± 0.15	0.31 ± 0.09
	Human	*0.08*	*0.33*	*0.83*

Table 2: Detailed analysis results per wave. *three decimal places, SynGemma outperforms $\text{ANT}_{\text{LLaMa}}$.

Sensitivity to Likert Scale Resolution. To assess whether our conclusions depend on fine-grained Likert calibration, we perform a sensitivity analysis using a coarser 3-point ordinal scale. Specifically, we collapse the original 5-point response categories into 3-point bins and recompute EMD, Frob., and Cron. α across the same evaluation settings. The qualitative ordering of methods remains largely unchanged under this coarser scale: $\text{SYN}_{\text{Gemma}}$ retains its lead in five of the six primary wave-level comparisons, and the correspondence between the 5-point and 3-point evaluations remains high across settings (Spearman $\rho \approx 0.87$; mean Pearson $r = 0.89$). These results suggest that the gains of $\text{SYN}_{\text{Gemma}}$ reflect robust improvements in opinion alignment rather than artifacts of fine-grained response calibration. Full results are provided in Appendix G.1.

Persona Consistency Analysis. To better understand the performance gap between SYNTHIA and purely synthetic baselines, we analyze internal factual consistency within persona narratives. Manual inspection revealed that Anthology personas frequently contain internally contradictory statements (Table 3). To systematically quantify this issue, we apply line-to-line inconsistency detection (Abdulhai et al., 2025) across all personas.

We use an LLM-based judge to identify inconsistent text spans within each persona. To ensure the reliability of this judge, we compare its outputs against human annotations on a subset of personas and benchmark three state-of-the-art API LLMs, selecting the model with the highest agreement

Source	Narratives with contradictions highlighted
Anthology	Growing up, I found solace in the magical worlds of Disney movies ... My love for these films began with classics like **'The Lion King' and 'Mary Poppins,'** which I watched with my parents. **These movies, released around the same time**, shared a similar vibe...
Synthia	I was **born into a wealthy family** in city X... met my wife in university studying psychology. **My parents were immigrants so I want to help them out with living expenses**. I dislike non-fiction books.

Table 3: Examples of narratives with inconsistencies. Bolded text indicates the contradicting statesments.

Dataset	% Personas with Contradiction	Mean Error per Persona
$\text{ANT}_{\text{LLaMa}}$	0.63	0.959
$\text{SYN}_{\text{Gemma}}$	0.18	0.221
$\text{PCHAT}_{\text{Human}}$	0.04	0.047

Table 4: Results of persona contradiction analysis.

# Posts Used	Faithfulness ($\mathcal{F}$)	Hallucination Rate ($\mathcal{H}$)
< 20	0.5762	0.4237
20–50	0.6244	0.3755
50–100	0.7078	0.2921
100–200	0.7518	0.2481
> 200	0.7534	0.2465

Table 5: Persona faithfulness to underlying social media posts across five ranges of post counts used during generation.

with human judgments (69%; see Table 7 and Appendix C). The resulting statistics are summarized in Table 4.

Using human-authored personas ($\text{PCHAT}_{\text{Human}}$) as a reference, we find that $\text{ANT}_{\text{LLaMa}}$ contains more than three times as many personas with at least one internal contradiction compared to $\text{SYN}_{\text{Gemma}}$ (0.63% vs. 0.18%). Because individual personas may contain multiple contradictory spans, we also compute the mean number of inconsistencies per persona. Under this metric, $\text{SYN}_{\text{Gemma}}$ again outperforms $\text{ANT}_{\text{LLaMa}}$ by a large margin, indicating substantially improved internal factual consistency.

Persona Faithfulness to Underlying Posts. To complement our analysis of internal consistency, we evaluate the faithfulness of generated personas to their underlying social media posts. Following established methods in factuality evaluation (Min et al., 2023; Laban et al., 2022; Honovich et al., 2021), we formulate this task as an atomic evaluation of factual precision over generated personas.

Specifically, we (1) decompose each generated persona into a set of atomic claims $\mathcal{C} = \{c_1, c_2, \ldots, c_N\}$, (2) retrieve the most relevant posts from the user's history for each claim, (3) apply an LLM-based entailment classifier to label each claim as *supported*, *contradicted*, or *unverifiable* given the retrieved evidence, and (4) compute faithfulness and hallucination rate, defined as $\mathcal{F} = S/N$ and $\mathcal{H} = (C + U)/N$, respectively, where $N = |\mathcal{C}|$ denotes the total number of atomic claims, S the number of supported claims, C the number of contradicted claims, and U the number of unverifiable claims. Note that $S + C + U = N$ by construction, so $\mathcal{F} + \mathcal{H} = 1$.

We conduct this experiment on a sample of 250 personas across five distinct ranges of post counts used during generation. To ensure the reliability of the automated evaluation, we manually inspect the LLM outputs at each stage of the pipeline and correct any errors.

The results (see Table 5) reveal a clear positive correlation between the number of posts used to generate a persona and its faithfulness to those posts. This finding suggests that providing the generator with more user content reduces the incidence of hallucinated claims.

5 Bias and Fairness Analysis

We present a comprehensive evaluation of fairness and bias in SYNTHIA relative to strong baselines. Our analysis examines subgroup-level fidelity, stability, and parity, assessing whether SYNTHIA personas represent diverse demographics with comparable accuracy and reliability. We ground this evaluation in established frameworks for algorithmic fairness (Barocas et al., 2023), focusing in particular on *representational harms*, where models may oversimplify minority groups or fail to capture within-group nuance.

Fidelity and Stability. We first analyze the relationship between fidelity and stability across demographic subgroups. Fidelity measures how closely synthetic personas reflect the opinions of their real-world counterparts, while stability captures the

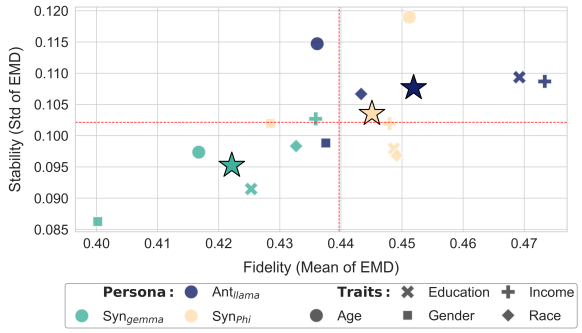


Figure 3: Relation between Fidelity and Stability (for both, lower is better). Red lines are the global average. Stars are the overall averages for each persona type.

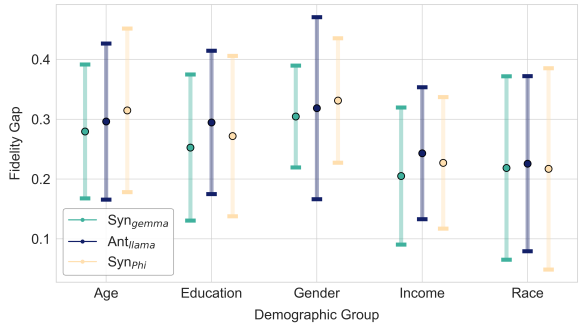


Figure 4: Subgroup Fidelity gap analysis. Values near zero indicate performance matching human variation.

consistency of this alignment across generations. Figure 3 visualizes this relationship. We quantify fidelity using the mean Earth Mover's Distance (EMD) per demographic group, and stability using the standard deviation of these EMD scores.

As shown in Figure 3, SYN$_{\text{Gemma}}$ consistently occupies the "Ideal" (bottom-left) quadrant. Averaged across all demographics, SYN$_{\text{Gemma}}$ attains a lower mean EMD than ANT$_{\text{LLaMa}}$, indicating improved fidelity. In addition, SYN$_{\text{Gemma}}$ exhibits lower variance across runs, reflecting greater stability. Together, these results show that while ANT$_{\text{LLaMa}}$ can achieve reasonable performance in some settings, SYN$_{\text{Gemma}}$ produces more reliable and consistently faithful representations across demographic groups.

Relative Fidelity and Human Baselines. Although raw EMD scores quantify distributional differences, they do not account for the inherent diversity and variance within real human populations across demographic subgroups. Groups with highly polarized opinions are intrinsically more difficult to simulate than those exhibiting broad consensus. To account for this, we normalize model

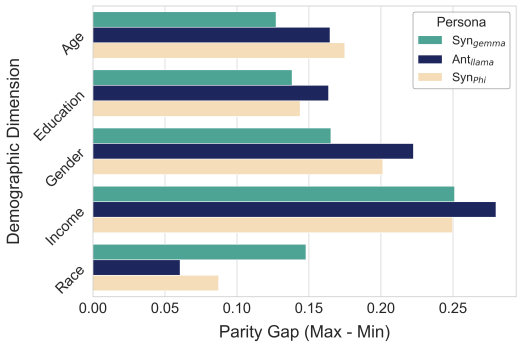


Figure 5: Parity gap analysis. Bars show the difference between best- and worst-simulated subgroups per demographic. Shorter bars indicate fairer treatment.

fidelity against real-world human opinion distributions. Specifically, we compute the *Fidelity Gap* as the difference between a model's EMD for a given subgroup and the internal EMD of the corresponding human subgroup, which serves as a lower bound on natural human variation (see Appendix D for implementation details).

Figure 4 shows the resulting relative fidelity gaps. SYN$_{\text{Gemma}}$ consistently exhibits smaller gaps across all demographics, with a lower average fidelity gap than ANT$_{\text{LLaMa}}$, indicating reduced divergence from natural human variation. This result demonstrates that SYNTHIA does not merely minimize distributional distance, but does so while respecting the underlying opinion diversity within demographic subgroups.

Parity Gap Analysis. Finally, we assess whether any subgroup within a demographic is systematically disadvantaged (e.g., whether high- and low-income groups are simulated with comparable accuracy). Large performance disparities across subgroups are a well-known indicator of algorithmic bias (Mehrabi et al., 2021). We define the *Parity Gap* as the difference between the maximum and minimum error observed among subgroups within each demographic category (see Appendix D.2 for details). Figure 5 reports the Parity Gap across demographics.

Our analysis shows that SYN$_{\text{Gemma}}$ achieves lower parity gaps in four out of five demographic categories. For example, whereas ANT$_{\text{LLaMa}}$ exhibits substantial disparities in Gender ($P_{\text{gap}} = 0.22$) and Income ($P_{\text{gap}} = 0.28$), SYN$_{\text{Gemma}}$ reduces these gaps to 0.17 and 0.25, corresponding to reductions of 22.7% and 10.7%, respectively. These results indicate a more equitable representation of demographic subgroups and highlight the

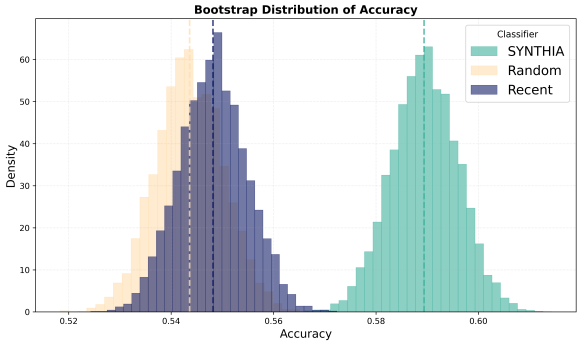


Figure 6: Bootstrap analysis of accuracy over G_{local} with different baselines (N = 10,000).

Model	Acc	Prec	Rec	F1
Panel A: Induced Random Subgraph (G_{global})				
SYNTHIA	**0.72**	**0.72**	0.70	**0.71**
Random Ext.	0.69	0.68	0.73	**0.71**
Recency Ext.	0.68	0.66	**0.74**	0.70
Panel B: Snowball Sampled Subgraph (G_{local})				
SYNTHIA	**0.59**	**0.56**	0.91	**0.69**
Random Ext.	0.54	0.53	0.93	0.67
Recency Ext.	0.55	0.53	**0.95**	0.68

Table 6: Link prediction performance on G_{global} and G_{local}. Best results in bold; Ext: extractive.

importance of reducing disparate impact when deploying synthetic personas for social science research, where representational harms must be carefully controlled.

6 Social Network Analysis

To evaluate the utility of SYNTHIA beyond survey simulation, we conduct two case studies on a real-world social network. We examine whether semantic similarity between personas correlates with network homophily among their corresponding nodes, assessing whether persona representations capture latent social signals aligned with observed follower–following relationships.

To capture both global and local structure, we extract two subgraphs from the full follower graph. The induced subgraph G_{global}, obtained via random node sampling, reflects global connectivity patterns, while the snowball-sampled subgraph G_{local} captures dense, community-level interactions. Sampling details are provided in Appendix E.

As this constitutes a new evaluation setting for SYNTHIA, we construct two length-matched extractive baselines from users' historical activity: a random baseline and a recency-based baseline that prioritizes recent posts.

Link Prediction. Our first case study evaluates whether persona text alone can predict real-world social connections. We fine-tune a transformer-based binary classifier to estimate the probability of an edge between two nodes using only the textual content of their associated personas (see Appendix E). By excluding all explicit graph features, this setup isolates the extent to which personas capture semantically grounded homophily aligned with network structure. This evaluation is designed to test whether generated personas encode latent social similarity rather than explicit graph cues: personas are generated from privacy-filtered source posts with direct identifiers and interaction markers removed.

Table 6 reports performance on both subgraphs. On G_{global}, SYNTHIA outperforms both extractive baselines in accuracy and F1 score ($p < 0.001$, McNemar's test; Panel A), indicating that its personas encode global patterns of social similarity more effectively than activity-based summaries. Performance on G_{local} is lower for all methods due to the dense, highly homophilous structure of the subgraph. The extractive baselines exhibit very high recall (> 0.93) but low precision (≈ 0.52), reflecting a degenerate strategy that largely defaults to predicting positive edges. This behavior indicates an inability to capture the fine-grained distinctions that separate actual follower relationships from general community membership.

In contrast, SYNTHIA demonstrates stronger discriminative capacity, achieving higher precision and F1 score while maintaining competitive recall. As shown in Panel B of Table 6, SYNTHIA again outperforms both baselines ($p < 0.001$), with a larger performance gap than in G_{global} (Figure 6). This suggests that SYNTHIA personas encode latent user interests and relational signals necessary to model fine-grained social structure.

These findings underscore that while extractive baselines prioritize coverage, evidenced by their near-universal recall on G_{local}, they fail to reliably distinguish true social ties from general community membership. By achieving superior precision compared to baselines, SYNTHIA effectively avoids this degenerate prediction behavior. The resulting F1 score highlights the model's robustness in dense settings where determining edge existence requires capturing subtle semantic affinities rather than simple activity recency. Consequently, the statistically

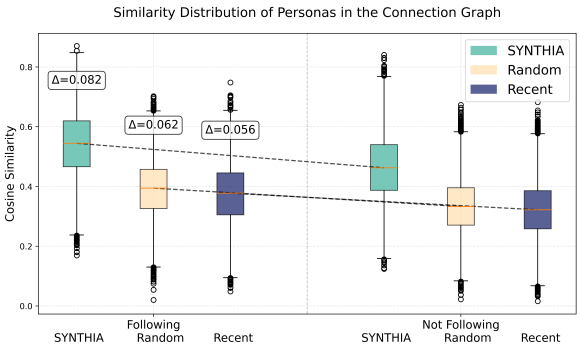


Figure 7: Semantic similarity gap between *following* and *not following* personas.

significant performance gains ($p < 0.001$) across both subgraphs confirm that our generated personas provide a more faithful representation of the underlying network homophily.

Similarity Distribution Analysis. To further examine the relationship between persona representations and network structure, we compute cosine similarity between persona embeddings for all node pairs with a follower relationship and an equally sized random sample of non-following pairs. This analysis is conducted for SYNTHIA and both extractive baselines. Figure 7 shows the resulting similarity distributions.

Across all methods, following pairs exhibit higher median similarity than non-following pairs, reflecting underlying social homophily. However, SYNTHIA demonstrates substantially stronger alignment with network structure. Its personas are more semantically cohesive, with a median cosine similarity of approximately 0.55 for following pairs, compared to 0.40 and 0.38 for the random and recency baselines, respectively. In addition, SYNTHIA shows the largest separation between following and non-following distributions, with a median difference of $\Delta = 0.082$, exceeding the random baseline by 32.2% ($\Delta = 0.062$) and the recency baseline by 46.4% ($\Delta = 0.056$).

These results indicate that SYNTHIA more effectively encodes the semantic signal of social connectivity into persona representations, aligning latent textual similarity with observed network homophily. This pattern is consistent across both global and local subgraphs, suggesting it is not driven by a particular sampling strategy. While this analysis does not assume a causal relationship between similarity and edge formation, it complements the link prediction results by providing a distributional view of how persona semantics correspond to social structure.

7 Conclusion

We introduced SYNTHIA, a persona synthesis framework that combines the scalability of large language models with grounding in real-world, human-generated data. By constructing personas from open social media content, SYNTHIA addresses key limitations of purely synthetic approaches, including internal narrative inconsistency and demographic bias.

Across extensive experimental settings, we show that SYNTHIA produces virtual populations that more faithfully align with population-level opinion distributions while exhibiting improved fairness and stability across sensitive demographic attributes. Moreover, by preserving the underlying social network structure of the source data, SYNTHIA enables the analysis of personas within realistic relational contexts, as demonstrated through our network-based case studies.

Looking ahead, this framework opens several avenues for future research. Virtual personas grounded in social networks can be used to study interventions aimed at reducing polarization or harmful content, as well as to model diverse interaction types while accounting for network structure and temporal dynamics. We emphasize that the generated personas reflect patterns present in the underlying data and should be interpreted accordingly. To support reproducibility and further empirical investigation, we release the personas together with their associated network metadata. We hope this work encourages more principled, transparent, and responsible use of persona-driven simulations in computational social science.

Limitations

SYNTHIA *should be interpreted as a tool for population and subgroup level simulation, not as a faithful reconstruction of any specific individual.* Parts of our evaluation pipeline still rely on LLM-based simulation and LLM-based judgment. Although our primary alignment metrics are anchored to human ATP distributions rather than judge scores, survey simulation may still inherit model-specific priors, prompt sensitivities, and calibration errors. Likewise, the narrative consistency analysis uses an LLM judge with imperfect agreement to human annotators. We mitigate these risks through a fac-

torial robustness design across surveyor models, randomized response order, use of base models for survey simulation, judge validation against human labels, and the additional faithfulness analysis introduced in this work. Still, these controls do not fully substitute for broader human-in-the-loop validation, especially for individual-level fidelity. While SYNTHIA establishes a robust framework for scalable and authentic persona generation, our current study highlights several avenues for future exploration. First, to ensure the reproducibility and accessibility of our pipeline for the broader academic community, we prioritized evaluations using open-weights models and consumer-grade hardware. While these models demonstrate high fidelity, extending the SYNTHIA framework to proprietary, frontier-class models could further enhance narrative nuance, though this remains outside the scope of the current open-science focus. Second, our validation of opinion alignment utilized the Bluesky social network due to its transparent data policies and open architecture. While this provides a rich and ethically sourced testbed, early-adopter communities on any single platform may exhibit specific sociodemographic distributions. We designed SYNTHIA to be platform-agnostic; thus, applying our methodology to alternative data sources in future work could capture an even broader spectrum of global demographic variances. Finally, our network analysis focused on static structural and semantic homophily to validate the integrity of the generated persona graph. Future research could dynamically model how these virtual personas evolve over time, offering deeper insights into temporal opinion shifts and longitudinal social dynamics.

Ethical Considerations

The development of high-fidelity virtual personas necessitates a rigorous commitment to data privacy and responsible AI usage, and we adhered to a strict ethical framework throughout the data lifecycle. All data was sourced exclusively from the Bluesky authenticated open data stream, strictly complying with the platform's terms of service and user redistribution policies. We processed only public-facing content, respecting the "right to be forgotten" by excluding protected or deleted accounts. To protect user anonymity, we implemented a multi-stage privacy pipeline prior to model ingestion. This involved utilizing automated Named Entity Recognition (NER) combined with heuristic filtering to detect and scrub Personally Identifiable Information (PII), including real names, physical addresses, and contact details from the source text. Furthermore, all original user identifiers were replaced with randomized, hashed tokens to ensure the released dataset contains no direct linkage to live social media profiles. We also obfuscated precise timestamps to prevent identification via temporal correlation attacks, retaining only the relative sequential order required to maintain narrative coherence. Finally, we acknowledge that generative agents are dual-use technologies. While SYNTHIA is designed for computational social science simulation, we explicitly prohibit its use for deceptive practices or manipulation. To this end, the released models and datasets are governed by a research-only license, and we have established a protocol to promptly remove any data points if future privacy concerns arise.

Acknowledgments

We acknowledge the use of AI assistance solely for grammar review and the generation of code necessary for producing plots and figures. Any AI-generated content represents a paraphrase of original material authored by the researchers, aimed at improving the readability of the text.

References

Marwa Abdulhai, Ryan Cheng, Donovan Clay, Tim Althoff, Sergey Levine, and Natasha Jaques. 2025. Consistently simulating human personas with multi-turn reinforcement learning. *Preprint*, arXiv:2511.00222.

Jacy Reese Anthis, Ryan Liu, Sean M. Richardson, Austin C. Kozlowski, Bernard Koch, James Evans, Erik Brynjolfsson, and Michael Bernstein. 2025. Llm social simulations are a promising research method. *Preprint*, arXiv:2504.02234.

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

Julia Barnett, Kimon Kieslich, and Nicholas Diakopoulos. 2024. Simulating policy impacts: Developing a generative scenario writing method to evaluate the perceived effects of regulation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 82–93.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press.

Ngoc Bui, Hieu Trung Nguyen, Shantanu Kumar, Julian Theodore, Weikang Qiu, Viet Anh Nguyen, and Rex Ying. 2025. Mixture-of-personas language models for population simulation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24761–24778, Vienna, Austria. Association for Computational Linguistics.

Aurélien Bück-Kaeffer, Je Qin Chooi, Dan Zhao, Maximilian Puelma Touzel, Kellin Pelrine, Jean-François Godbout, Reihaneh Rabbany, and Zachary Yang. 2025. BluePrint: A social media user dataset for llm persona evaluation and training. *Preprint*, arXiv:2510.02343.

Hongzhan Chen, Hehong Chen, Ming Yan, Wenshen Xu, Gao Xing, Weizhou Shen, Xiaojun Quan, Chenliang Li, Ji Zhang, and Fei Huang. 2024a. SocialBench: Sociality evaluation of role-playing conversational agents. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2108–2126, Bangkok, Thailand. Association for Computational Linguistics.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, Aili Chen, Nianqi Li, Lida Chen, Caiyu Hu, Siye Wu, Scott Ren, Ziquan Fu, and Yanghua Xiao. 2024b. From persona to personalization: A survey on role-playing language agents. *Preprint*, arXiv:2404.18231.

Myra Cheng, Tiziano Piccardi, and Diyi Yang. 2023. CoMPosT: Characterizing and evaluating caricature in LLM simulations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10853–10875, Singapore. Association for Computational Linguistics.

Hyeong Kyu Choi and Yixuan Li. 2024. Picle: Eliciting diverse behaviors from large language models with persona in-context learning. In *International Conference on Machine Learning*, pages 8722–8739. PMLR.

Celia Cintas, Miriam Rateike, Erik Miehling, Elizabeth Daly, and Skyler Speakman. 2025. Localizing persona representations in llms. *Preprint*, arXiv:2505.24539.

Tejaswani Dash, Dinesh Karri, Anudeep Vurity, Gautam Datla, Tazeem Ahmad, Saima Rafi, and Rohith Tangudu. 2025. Polypersona: Persona-grounded llm for synthetic survey responses. *Preprint*, arXiv:2512.14562.

Xiangyang He, Jiale Li, Jiahao Chen, Yang Yang, and Mingming Fan. 2025. Simupanel: A novel immersive multi-agent system to simulate interactive expert panel discussion. *Preprint*, arXiv:2506.16010.

Zihao He, Minh Duc Chu, Rebecca Dorn, Siyi Guo, and Kristina Lerman. 2024. Community-cross-instruct: Unsupervised instruction generation for aligning large language models to online communities. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17001–17019, Miami, Florida, USA. Association for Computational Linguistics.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. q^2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7856–7870, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Soon-Gyo Jung, Joni Salminen, Kholoud Khalil Aldous, and Bernard J. Jansen. 2025. Personacraft: Leveraging language models for data-driven persona development. *International Journal of Human-Computer Studies*, 197:103445.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Ang Li, Haozhe Chen, Hongseok Namkoong, and Tianyi Peng. 2025. Llm generated persona is a promise with a catch. *Preprint*, arXiv:2503.16527.

Andy Liu, Mona Diab, and Daniel Fried. 2024. Evaluating large language model biases in persona-steered generation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9832–9850, Bangkok, Thailand. Association for Computational Linguistics.

Genglin Liu, Vivian T. Le, Salman Rahman, Elisa Kreiss, Marzyeh Ghassemi, and Saadia Gabriel. 2025. MOSAIC: Modeling social AI for content dissemination and regulation in multi-agent simulations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6401–6428, Suzhou, China. Association for Computational Linguistics.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural*

Language Processing, pages 12076–12100, Singapore. Association for Computational Linguistics.

Suhong Moon, Marwa Abdulhai, Minwoo Kang, Joseph Suh, Widyadewi Soedarmadji, Eran Kohen Behar, and David Chan. 2024. Virtual personas for language models via an anthology of backstories. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19864–19897, Miami, Florida, USA. Association for Computational Linguistics.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

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

Pew Research Center. 2018. American trends panel wave 34. Dataset. Field dates: April 23 – May 6, 2018. Topics: Biomedical and food issues.

Pew Research Center. 2021. American trends panel wave 99. Dataset. Field dates: Nov. 1 – Nov. 7, 2021. Topics: Artificial Intelligence (AI) and human enhancement.

Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, Chen Gao, Fengli Xu, Fang Zhang, Ke Rong, Jun Su, and Yong Li. 2025. Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of human behaviors and society. *Preprint*, arXiv:2502.08691.

Giorgio Piatti, Zhijing Jin, Max Kleiman-Weiner, Bernhard Schölkopf, Mrinmaya Sachan, and Rada Mihalcea. 2024. Cooperate or collapse: Emergence of sustainable cooperation in a society of llm agents. *Advances in Neural Information Processing Systems*, 37:111715–111759.

Nusrat Jahan Prottasha, Md Kowsher, Hafijur Raman, Israt Jahan Anny, Prakash Bhat, Ivan Garibay, and Ozlem Garibay. 2025. User profile with large language models: Construction, updating, and benchmarking. *Preprint*, arXiv:2502.10660.

Vahid Rahimzadeh, Ali Hamzehpour, Azadeh Shakery, and Masoud Asadpour. 2025. From millions of tweets to actionable insights: Leveraging llms for user profiling. *arXiv preprint arXiv:2505.06184*.

Leonard Salewski, Stephan Alaniz, Isabel Rio-Torto, Eric Schulz, and Zeynep Akata. 2023. In-context impersonation reveals large language models' strengths and biases. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Jocelyn J Shen, Akhila Yerukola, Xuhui Zhou, Cynthia Breazeal, Maarten Sap, and Hae Won Park. 2025. Words like knives: Backstory-personalized modeling and detection of violent communication. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 11607–11625, Suzhou, China. Association for Computational Linguistics.

Christine V Stephens and Mary Breheny. 2013. Narrative analysis in psychological research: An integrated approach to interpreting stories. *Qualitative Research in Psychology*, 10:14 – 27.

Joseph Suh, Erfan Jahanparast, Suhong Moon, Minwoo Kang, and Serina Chang. 2025a. Language model fine-tuning on scaled survey data for predicting distributions of public opinions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21147–21170, Vienna, Austria. Association for Computational Linguistics.

Joseph Suh, Suhong Moon, and Serina Chang. 2025b. Rethinking llm human simulation: When a graph is what you need. *Preprint*, arXiv:2511.02135.

Lipeipei Sun, Tianzi Qin, Anran Hu, Jiale Zhang, Shuojia Lin, Jianyan Chen, Mona Ali, and Mirjana Prpa. 2025. Persona-l has entered the chat: Leveraging llms and ability-based framework for personas of people with complex needs. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025a. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Microsoft Team, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, Xiyang Dai, and 56 others. 2025b. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *Preprint*, arXiv:2503.01743.

Maximilian Puelma Touzel, Sneheel Sarangi, Austin Welch, Gayatri Krishnakumar, Dan Zhao, Zachary Yang, Hao Yu, Ethan Kosak-Hine, Tom Gibbs, Andreea Musulan, Camille Thibault, Busra Tugce Gurbuz, Reihaneh Rabbany, Jean-François Godbout,

and Kellin Pelrine. 2024. A simulation system towards solving societal-scale manipulation. *Preprint*, arXiv:2410.13915.

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

Kuang Wang, Xianfei Li, Shenghao Yang, Li Zhou, Feng Jiang, and Haizhou Li. 2025a. Know you first and be you better: Modeling human-like user simulators via implicit profiles. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21082–21107, Vienna, Austria. Association for Computational Linguistics.

Lei Wang, Jingsen Zhang, Hao Yang, Zhi-Yuan Chen, Jiakai Tang, Zeyu Zhang, Xu Chen, Yankai Lin, Hao Sun, Ruihua Song, Xin Zhao, Jun Xu, Zhicheng Dou, Jun Wang, and Ji-Rong Wen. 2025b. User behavior simulation with large language model-based agents. *ACM Trans. Inf. Syst.*, 43(2).

Dulce Wilkinson Westberg, Moin Syed, Aerika Brittian Loyd, and William Dunlop. 2024. Using intersectionality to understand how structural domains are embedded in life narratives. *Journal of personality*.

Rui Xu, Xintao Wang, Jiangjie Chen, Siyu Yuan, Xinfeng Yuan, Jiaqing Liang, Zulong Chen, Xiaoqing Dong, and Yanghua Xiao. 2024. Character is destiny: Can role-playing language agents make persona-driven decisions? *Preprint*, arXiv:2404.12138.

Min Yin, Haoyu Liu, Boyi Lian, and Chunlei Chai. 2025. Co-persona: Leveraging llms and expert collaboration to understand user personas through social media data analysis. *Preprint*, arXiv:2506.18269.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

A Experimental Setups

A.1 Models

We employed various language models for different components of our pipeline:

- **Llama 3 8B & Gemma 27B & 12B**: Used for the demographic surveying and ATP question answering components. Gemma 27B is also utilized for persona generation.
- **Phi-4-mini-instruct 4B**: Utilized for both persona generation from social network history and response parsing in the demographic surveying phase.
- **Gemini 2.5 Flash & Claude 3.7 Sonnet & GPT-4.1**: Used for the inconsistency detection pipeline, accessed through the available APIs on Openrouter.ai platform.
- **ModernBERT-base & Qwen3-Embedding-0.6B**: Used for the link prediction and distribution analysis respectively.

Model	Human Agreement (%)
Gemini 2.5 Flash	69
Claude 3.7 Sonnet	61
GPT-4.1	56

Table 7: Human agreement rates across different language models.

A.2 Hardware and Deployment

All models, except Gemma-27B, were served on two RTX 6000 GPUs and one RTX 8000 GPU. The RTX 6000 machines each have 24 GB of VRAM and 128 GB of system RAM, while the RTX 8000 machine has 48 GB of VRAM and 256 GB of system RAM. Gemma-27B was served on Vertex AI using two H100 GPUs with 160 GB of VRAM.

A.3 Software

LLaMA 3 8B, Gemma 27B and 12B, and Phi-4 Mini Instruct were all served using vLLM with Python 3.10. The CUDA version used across all GPUs was 12.4. For the embedding model, the Sentence-Transformers and Transformers libraries were used.

A.4 Hyperparameters

- For demographic surveying and ATP question answering, we used the default hyperparameters specified in the original Anthology paper.
- For backstory generation, we set the temperature to 0.1 and limited the maximum number of generated tokens to 1500.
- We used 42 as the random seed for case studies in both model initialization and dataset splitting.

- To replicate Anthology, we adopted all hyperparameters reported in their paper and GitHub repository.

B Dataset Generation

We clean each of users and corresponding posts by de-duplication, removing posts with unusual dates (such as 1/1/1), and removing non-English posts. We used "langdetect" library to label each post with a language.

Unless otherwise specified, $\text{SYN}_{\text{Gemma}}$ personas are generated with Gemma-3-27B and SYN_{Phi} personas with Phi-4-mini-instruct, with the prompt illustrated in Figure 9. A sample of personas with grounding data is presented in Figure 8.

Social Media Data Statistics			
Total posts			667,842
Mean posts per user			234.50 ± 262.88
Mean words per post			15.17 ± 13.97
Max words per post			100
Persona Data Statistics (Words per Persona)			
Dataset	**Min**	**Max**	**Mean**
$\text{SYN}_{\text{Gemma}}$	194	559	310.83
SYN_{Phi}	121	351	257.48
$\text{ANT}_{\text{LLaMa}}$	8	1,376	376.56
$\text{ANT}_{\text{Gemma}}$	21	1,354	506.67
$\text{PCHAT}_{\text{Human}}$	14	76	32.51
Social Network Metrics (Pruned Graph)			
Metric			**Value**
# nodes (N)			2,053
# edges (M)			26,467
Network density			0.0063
Mean clustering coefficient			0.2695
Connectivity (Pruned Graph)			
Weakly connected components			11
Strongly connected components			502
Largest WCC size			2,031

Table 8: Synthesized data created by SYNTHIA used in this work. Network statistics are computed on the pruned graph after removing isolated nodes.

C Consistency Analysis

Consistency analysis was conducted with three different API state of the art LLMs and their results were compared to that of humans. "google/gemini-2.5-flash" API had the most corelation with human annotators (see Table 7), therefore we used this model as our judge to detect contradictions across different sets of personas. All the models were accessed through OpenRouter[2]. Prompt template illustrated in Figure 10. The said model respected the output format for nearly all the cases and the following regex pattern used to parse the outputted JSON from the Judge model:

```
```json\s*(.*?)\s*```
```

D Bias Analysis

This section details the mathematical formulations used to evaluate the fairness and fidelity of the SYNTHIA personas in Section 5. We introduce two key metrics: the *Fidelity Gap*, which measures accuracy relative to natural human variance, and the *Parity Gap*, which quantifies the disparity in performance across different demographic subgroups.

D.1 Fidelity Gap

Standard distance metrics like Earth Mover's Distance (EMD) can be misleading if they do not account for the inherent diversity within a target population. A subgroup with high internal disagreement (e.g., a "purple" state in political polling) is naturally harder to simulate than a homogenous group. To address this, we define the **Fidelity Gap** as the excess error of the model over the natural internal variation of the human population.

For a demographic category D (e.g., Gender, Age, Race), we first compute the gap for each subgroup $d \in D$, then average across all subgroups as shown in Equation 1:

$$\text{Fidelity Gap}_D = \frac{1}{|D|} \sum_{d \in D} \Big[\text{EMD}(P^d_{\text{human}}, P^d_{\text{LLM}}) - \text{EMD}^{\text{internal},d}_{\text{human}} \Big] \tag{1}$$

The critical component here is the *Internal Human EMD* ($\text{EMD}^{\text{internal},d}_{\text{human}}$), which serves as a baseline for the "irreducible" variance within a group. This is calculated via a bootstrapping approach:

$$\text{EMD}^{\text{internal},d}_{\text{human}} = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{|Q|} \sum_{q \in Q} W_1(P^{d,(k,1)}_{\text{human},q}, P^{d,(k,2)}_{\text{human},q}) \tag{2}$$

[2] http://openrouter.ai/
```

Social Media Posts: "pls just ghost me. each rejection email hurts me" gifted tarot deck to me, probably already used... my cat fell asleep on my chest ... been together almost 5 years. "feel better after getting back into my spirituality." "if i get christmas money, im finally ordering from this amazing page... 🙏" "#art #artsky #patterns #print #colors #harmony"

Generated backstory with color matching between narrative part corresponding to social media post:
My journey began...always fascinated by the interplay of colors and patterns...My love for art took a significant turn when I stumbled upon a local art scene that celebrated creativity and expression .... However, the road to success was not without its challenges. Rejection became a familiar sting, each email a reminder of the hurdles I had to overcome....My life took an unexpected turn when I discovered a used tarot deck...I found solace in my spirituality...My cat, a loyal companion for almost five years, became a symbol of comfort and joy in my life. Despite the challenges, I remained grateful for the small blessings, like the Christmas money that allowed me to finally order from my favorite art store...

Figure 8: Illustrative example from SYNTHIA showing a persona and its grounding social media posts. Highlights demonstrate how different spans of the persona relate to their respective source posts.

```
bs_gen_prompt_template = '''You are an expert creative writer. Based on the following posts from a social media
user, create a detailed backstory (at least two paragraphs) in first-person view that explains this person's life
journey, key events that shaped them, and their motivations.
User Posts:
{POSTS}

Place your answer in this JSON schema:
{{
    "post_ids": <List[int], a list of post ids that are significant for the backstory>,
    "backstory": <str, the generated backstory from the selected significant post ids, do not cite or mention the
ids in the generated backstory>
}}
'''
```

Figure 9: Prompt template for backstory generator model.

```
prompt_template = '''Read the following story. Give me a list of text span pairs that are visibly contradict each
other.
If you did not find any contradiction, return an empty list.
{BACKSTORY}
'''
```

Figure 10: Prompt template for inconsistency detector model.

where:

- D is a demographic category (e.g., gender = {male, female})
- $d \in D$ is a specific subgroup within that category (e.g., male)
- P^d_{human} is the weighted response distribution of humans in subgroup d
- P^d_{LLM} is the response distribution of LLM personas assigned to subgroup d
- $\text{EMD}^{\text{internal},d}_{\text{human}}$ represents the lower-bound proxy for natural human variance within subgroup d
- K is the number of random splits (we use $K = 20$)
- Q is the set of all questions of interest
- W_1 denotes the Wasserstein-1 distance (Earth Mover's Distance)
- $P^{d,(k,i)}_{\text{human},q}$ is the weighted response distribution for question q among humans in subgroup d in the i-th half of the k-th random split

To compute $\text{EMD}^{\text{internal},d}_{\text{human}}$ for each subgroup d, we randomly split the human population within that subgroup into two equal halves K times. For each split, we calculate the EMD between the two halves across all questions, then average over all splits and questions. The final Fidelity Gap for demographic category D is the average of the individual subgroup gaps.

D.2 Parity Gap

To ensure that the model does not disproportionately fail for specific minority or marginalized groups, we utilize the **Parity Gap** (P_D). This metric captures the "worst-case" disparity in simulation quality within a demographic category.

Formally, for a given demographic D containing a set of subgroups S, let EMD_s be the Earth Mover's Distance between the human and model distributions for subgroup $s \in S$. The Parity Gap is defined as the range between the best-performing and worst-performing subgroups:

$$P_D = \max_{s \in S}(EMD_s) - \min_{s \in S}(EMD_s) \quad (3)$$

A lower P_D indicates a more equitable model where fidelity is consistent regardless of the specific subgroup (e.g., the model simulates high-income and low-income individuals with comparable accuracy), minimizing representational harms.

E Social Network Analysis Details

This appendix provides the technical specifications for the graph sampling strategies, baseline construction, and the link prediction methodology used in Section 6.

E.1 Graph Sampling Strategy

To capture different topological properties of the social network, we extract two distinct test subgraphs from the primary ground-truth graph $G = (V, E)$.

Global Subgraph (G_{global}). Also referred to as G_{rand}, this subgraph is constructed via uniform random node sampling to represent the global network structure. We define a subset of test nodes $V_{\text{global}} \subset V$ by selecting 10% of the total nodes uniformly at random. The associated test edge set is defined as all directed edges in E originating from these nodes:

$$E_{\text{global}} = \{(u, v) \in E \mid u \in V_{\text{global}}\}$$

Local Subgraph (G_{local}). Also referred to as G_{conn}, this subgraph is designed to test the model's performance in dense, high-homophily neighborhoods. We select a subset of nodes $V_{\text{local}} \subset V$, where $|V_{\text{local}}| \approx 0.1|V|$, such that V_{local} forms a single connected component within the directed graph. This snowball sampling approach ensures that the test cases represent a localized community structure rather than disparate actors.

E.2 Baseline Specifications

To isolate the impact of SYNTHIA's generative approach, we compare against two extractive baselines derived from users' historical activity. Crucially, both baselines are constrained by a length-matching parameter to ensure parity with the synthesized personas. Let $L(P_{\text{syn}})$ be the token length of the synthesized persona for a given user.

Random Extractive Baseline (B_{rand}): For each user, we perform random sampling without replacement from their chronological post history. Posts are aggregated until the total character length $L(B_{\text{rand}})$ approximately matches $L(P_{\text{syn}})$. This baseline captures the user's "average" historical signal without recency bias.

Recency-Based Baseline (B_{rec}): We select the user's most recent posts in descending chronological order. This "latest-first" aggregation continues until $L(B_{\text{rec}}) \approx L(P_{\text{syn}})$, capturing the user's most contemporary linguistic patterns and temporal interests.

E.3 Link Prediction Methodology

We formulate link prediction as a binary classification task to evaluate the information density of the personas. Given two personas v_i and v_j, the model learns the conditional probability of an edge existence $(v_i, v_j) \in E$.

We fine-tune a transformer-based encoder (ModernBERT-base) to minimize the cross-entropy loss for the probability:

$$P((v_i, v_j) \in E \mid \mathbf{x}) = \sigma(\mathbf{W} \cdot \text{enc}(\text{concat}(v_i, v_j)) + b) \quad (4)$$

where $\text{enc}(\cdot)$ is the transformer encoder output, $\text{concat}(v_i, v_j)$ represents the concatenated text of the two personas, and σ is the sigmoid function.

Statistical Evaluation. We report metrics (Accuracy, Precision, Recall, F1) alongside 95% bootstrap confidence intervals calculated over 10,000 iterations. To determine statistical significance between SYNTHIA and the baselines, we utilize McNemar's test with continuity correction ($\alpha = 0.05$), which is appropriate for comparing paired binary classification results.

F ATP Details

F.1 Demographic Matching Algorithm

Demographic matching, proposed by (Moon et al., 2024), is an algorithm that identifies the closest persona/backstory to a human by comparing demographic traits. This algorithm samples a subpopulation from our backstory database that best demographically represents the human population for an ATP survey. The algorithm creates a bipartite graph where each backstory and real human is represented by a vertex, with edges representing their demographic similarity.

Here is a formal description of the algorithm:

Let vertex set $H = \{h_1, h_2, ..., h_n\}$ represent a set of n humans, while vertex set $V = \{v_1, v_2, ..., v_m\}$ represents a set of m backstories. Each human $h_i = (t_{i1}, t_{i2}, ..., t_{ik})$ consists of k demographic traits, and each backstory $v_j = (P(d_{j1}), P(d_{j2}), ..., P(d_{jk}))$ represents a probability distribution of demographic traits. The edge $e_{ij} \in E$ connects human h_i and backstory v_j.

The weight of edge $w(e_{ij})$ is defined as the product of likelihoods that the j-th backstory's traits correspond to the demographic traits of the i-th real human. Formally:

$$w(e_{ij}) = w(h_i, v_j) = \prod_{l=1}^{k} P(d_{jl} = t_{il})$$

The demographic matching can then be defined as the following optimization problem:

$$\pi : [n] \to [m]$$

```
Question: What is your age?
(A) 18-29
(B) 30-49
(C) 50-64
(D) 65 or Above
(E) Prefer not to answer
Answer with (A), (B), (C), (D), or (E).
Answer:
```

Figure 11: Prompt for demographic trait question: age

```
Question: What is your gender?
(A) Male
(B) Female
(C) Other (e.g., non-binary, trans)
(D) Prefer not to answer
Answer with (A), (B), (C), or (D).
Answer:
```

Figure 12: Prompt for demographic trait question: gender

$$\pi^* = \arg\max_{\pi} \sum_{i=1}^{n} w(h_i, v_{\pi(i)})$$

We implement a greedy matching approach, where it is not required to match each backstory to exactly one human (i.e., humans can share backstories).

F.2 Demographic Traits

In total we have five demographic traits. For each of these a question has been created and asked a non-instruct LLM to answer it 40 times. See Figure 11 for age, Figure 12 for gender, Figure 13 for education, Figure 14 for income, and Figure 15 for race and ethnicity.

F.3 Waves

Below are the questions of each ATP waves used in this study. The prompt templates are those used in the Anthology (Moon et al., 2024). Below is the exact questions and options used for surveying ATP waves in simulations.

F.3.1 Wave 34

- **Affordability of GMOs:** "How likely is it that genetically modified foods will lead to more affordably-priced food?"
  - Very likely
  - Fairly likely
  - Not too likely
  - Not at all likely
- **Health Problems from GMOs:** "How likely is it that genetically modified foods will lead to health problems for the population as a whole?"
  - Very likely
  - Fairly likely
  - Not too likely
  - Not at all likely
- **Environmental Impact of GMOs:** "How likely is it that genetically modified foods will create problems for the environment?"
  - Very likely
  - Fairly likely
  - Not too likely
  - Not at all likely
- **Personal Concern (GMOs):** "How much do you, personally, care about the issue of genetically modified foods?"
  - A great deal
  - Some
  - Not too much
  - Not at all
- **Organic Consumption:** "How much of the food you eat is organic?"

```
Question: What is the highest level of education you have
completed?
(A) Less than high school
(B) High school graduate or equivalent (e.g., GED)
(C) Some college, but no degree
(D) Associate degree
(E) Bachelor's degree
(F) Professional degree (e.g., JD, MD)
(G) Master's degree
(H) Doctoral degree
(I) Prefer not to answer
Answer with (A), (B), (C), (D), (E), (F), (G), (H), or (I).
Answer:
```

Figure 13: Prompt for demographic trait question: education

```
Question: What is your annual household income?
(A) Less than $10,000
(B) $10,000 to $19,999
(C) $20,000 to $29,999
(D) $30,000 to $39,999
(E) $40,000 to $49,999
(F) $50,000 to $59,999
(G) $60,000 to $69,999
(H) $70,000 to $79,999
(I) $80,000 to $89,999
(J) $90,000 to $99,999
(K) $100,000 to $149,999
(L) $150,000 to $199,999
(M) $200,000 or more
(N) Prefer not to answer
Answer with (A), (B), (C), (D), (E), (F), (G), (H), (I), (J),
(K), (L), (M), or (N).
Answer:
```

Figure 14: Prompt for demographic trait question: income

```
Question: Which of the following racial or ethnic groups do
you identify with?
(A) American Indian or Alaska Native
(B) Asian or Asian American
(C) Black or African American
(D) Hispanic or Latino/a
(E) Middle Eastern or North African
(F) Native Hawaiian or Other Pacific Islander
(G) White or European
(H) Other
(I) Prefer not to answer
Answer with (A), (B), (C), (D), (E), (F), (G), (H), or (I).
Answer:
```

Figure 15: Prompt for demographic trait question: race and ethnicity

  - Most of it
  - Some of it
  - Not too much
  - None at all

- **Antibiotics and Hormones:** "How much health risk, if any, does eating meat from animals that have been given antibiotics or hormones have for the average person over the course of their lifetime?"
  - A great deal of health risk
  - Some health risk
  - Not too much health risk
  - No health risk at all

- **Artificial Coloring:** "How much health risk, if any, does eating food and drinks with artificial coloring have for the average person over the course of their lifetime?"
  - A great deal of health risk
  - Some health risk
  - Not too much health risk
  - No health risk at all

- **Artificial Preservatives:** "How much health risk, if any, does eating food and drinks with artificial preservatives have for the average person over the course of their lifetime?"
  - A great deal of health risk
  - Some health risk
  - Not too much health risk
  - No health risk at all

F.3.2 Wave 99

- **AI Knowing Thoughts and Behaviors:** "How excited or concerned would you be if artificial intelligence computer programs could know people's thoughts and behaviors?"
  - Very excited
  - Somewhat excited
  - Equal excitement and concern
  - Somewhat concerned
  - Very concerned

- **AI Performing Household Chores:** "How excited or concerned would you be if artificial intelligence computer programs could perform household chores?"
  - Very excited
  - Somewhat excited
  - Equal excitement and concern
  - Somewhat concerned
  - Very concerned

- **AI Making Important Life Decisions:** "How excited or concerned would you be if artificial intelligence computer programs could make important life decisions for people?"
  - Very excited
  - Somewhat excited
  - Equal excitement and concern
  - Somewhat concerned
  - Very concerned

- **AI Diagnosing Medical Problems:** "How excited or concerned would you be if artificial intelligence computer programs could diagnose medical problems?"
  - Very excited
  - Somewhat excited
  - Equal excitement and concern
  - Somewhat concerned
  - Very concerned

- **AI Performing Repetitive Workplace Tasks:** "How excited or concerned would you be if artificial intelligence computer programs could perform repetitive workplace tasks?"
  - Very excited
  - Somewhat excited
  - Equal excitement and concern
  - Somewhat concerned
  - Very concerned

- **AI Handling Customer Service:** "How excited or concerned would you be if artificial intelligence computer programs could handle customer service calls?"

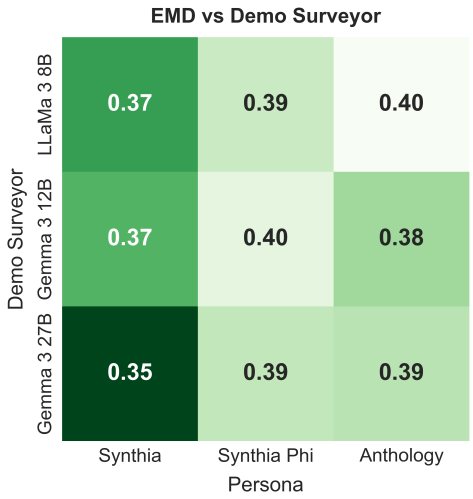


Figure 16: EMD per demo surveyor. Lower is better.

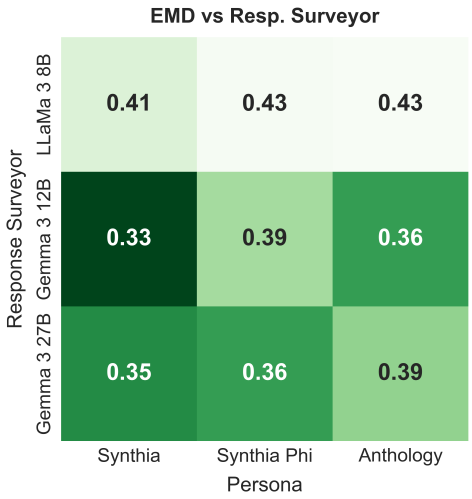


Figure 17: EMD per response surveyor. Lower is better.

- Very excited
- Somewhat excited
- Equal excitement and concern
- Somewhat concerned
- Very concerned

G Opinion Alignment Full Results

This section presents the comprehensive quantitative results for our opinion alignment analysis. We report Earth Mover's Distance (EMD), Frobenius norm, and Cronbach's α to evaluate the alignment accuracy and internal consistency of the surveyor models. Tables 10 and 11 provide a granular breakdown of interaction dynamics for Wave 34 and Wave 99, respectively, differentiating between demographic and response surveyors across standard backstories. Table 12 summarizes these findings with an aggregated performance view across all waves. Finally, Table 9 offers the full results for screening stage. The heatmaps on the relation between Cronbach Alpha and Frobenius Norm metrics with various models as demographic surveyor or response surveyor are given in Figure 18 and Figure 19

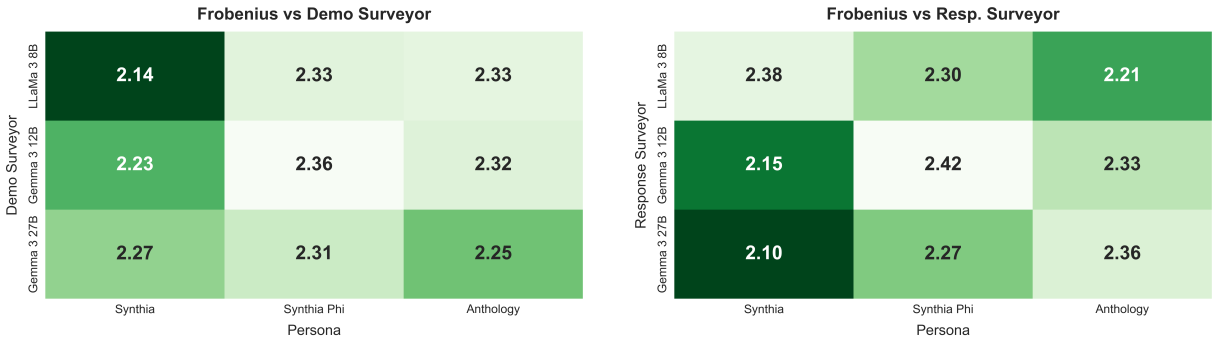
G.1 Sensitivity to Likert Scale Resolution

To assess whether the relative performance of persona sets depends on the granularity of the original Likert scale, we perform a sensitivity analysis using a coarser 3-point ordinal formulation. For each survey item, we collapse the original 5-point response categories into three bins while preserving the ordinal structure of the question, and recompute EMD, Frob., and Cron. α across the same evaluation settings used in the main analysis.

Table 9: Gemma 12 Performance Comparison across Waves 34 and 99

Wave	Backstory	EMD ↓	Frob. ↓	Cron. α ↑
34	ant	**0.341**	2.407	0.348
	ant_g27	0.341	2.649	0.315
	bsky	0.362	**2.246**	**0.381**
	bsky_phi	0.382	2.607	0.308
	PCHAT$_{\text{Human}}$	0.350	2.760	0.285
99	ant	0.372	**2.025**	**0.412**
	ant_g27	0.486	2.412	0.197
	bsky	**0.338**	2.210	0.337
	bsky_phi	0.449	2.393	0.167
	PCHAT$_{\text{Human}}$	0.578	2.048	0.380

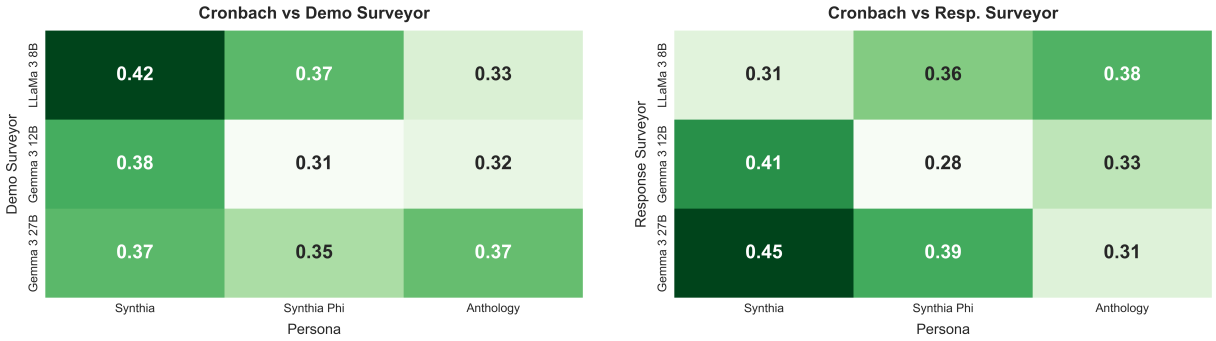
As shown in Table 13, the qualitative ranking of methods remains stable under this coarser scale. In particular, SYN$_{\text{Gemma}}$ remains the strongest overall system in five of the six primary wave-level comparisons. We further quantify agreement between the 5-point and 3-point evaluations in Table 14, observing high correspondence across settings. These results indicate that the improvements of SYN$_{\text{Gemma}}$ are robust to response-scale resolution and are not driven by fine-grained calibration artifacts.



(a) Frobenius Norm by Demographic Surveyor

(b) Frobenius Norm by Response Surveyor

Figure 18: Frobenius Norm performance heatmaps. Left: Aggregated by Demographic Surveyor. Right: Aggregated by Response Surveyor. Lower values indicate better alignment with the ground-truth correlation matrix.



(a) Cronbach's α by Demographic Surveyor

(b) Cronbach's α by Response Surveyor

Figure 19: Cronbach's α reliability heatmaps. Left: Aggregated by Demographic Surveyor. Right: Aggregated by Response Surveyor. Higher values indicate greater internal consistency.

Table 10: Wave 34: Detailed Interaction Analysis (Ant, Bsky, Bsky-Phi)

Wave	Demo Surveyor	Resp. Surveyor	Backstory	EMD ↓	Frob. ↓	Cron. α ↑
34	gemma_12	gemma_12	ant	**0.341**	2.407	0.348
			bsky	0.362	**2.246**	**0.381**
			bsky_phi	0.382	2.607	0.308
		gemma_27	ant	**0.325**	2.474	0.378
			bsky	0.335	**2.162**	**0.482**
			bsky_phi	0.332	2.327	0.426
		llama8	ant	**0.359**	2.449	0.345
			bsky	0.402	2.499	0.233
			bsky_phi	0.425	**2.425**	**0.370**
34	gemma_27	gemma_12	ant	**0.349**	2.506	0.239
			bsky	0.349	**2.229**	**0.430**
			bsky_phi	0.394	2.352	0.416
		gemma_27	ant	0.344	2.519	0.332
			bsky	**0.294**	**2.357**	**0.461**
			bsky_phi	0.347	2.426	0.386
		llama8	ant	0.385	**2.447**	**0.379**
			bsky	**0.366**	2.532	0.291
			bsky_phi	0.402	2.474	0.326
34	llama8	gemma_12	ant	0.351	2.540	0.289
			bsky	**0.319**	**2.096**	**0.457**
			bsky_phi	0.352	2.500	0.306
		gemma_27	ant	0.379	2.384	0.349
			bsky	**0.337**	**2.043**	0.454
			bsky_phi	0.347	2.286	**0.477**
		llama8	ant	**0.353**	**2.380**	0.361
			bsky	0.385	2.577	0.333
			bsky_phi	0.439	2.492	**0.392**

Table 11: Wave 99: Detailed Interaction Analysis (Ant, Bsky, Bsky-Phi)

Wave	Demo Surveyor	Resp. Surveyor	Backstory	EMD ↓	Frob. ↓	Cron. α ↑
99	gemma_12	gemma_12	ant	0.372	**2.025**	**0.412**
			bsky	**0.338**	2.210	0.337
			bsky_phi	0.449	2.393	0.167
		gemma_27	ant	0.438	2.442	0.136
			bsky	0.404	**2.041**	**0.441**
			bsky_phi	**0.352**	2.212	0.301
		llama8	ant	0.467	**2.138**	0.337
			bsky	**0.404**	2.219	**0.386**
			bsky_phi	0.445	2.176	0.317
99	gemma_27	gemma_12	ant	0.390	2.158	**0.378**
			bsky	**0.286**	**2.153**	0.353
			bsky_phi	0.400	2.193	0.307
		gemma_27	ant	0.390	**2.074**	**0.424**
			bsky	0.395	2.187	0.329
			bsky_phi	**0.360**	2.118	0.348
		llama8	ant	0.482	**1.731**	**0.467**
			bsky	**0.427**	2.123	0.357
			bsky_phi	0.445	2.215	0.290
99	llama8	gemma_12	ant	0.367	2.305	0.326
			bsky	**0.312**	**1.984**	**0.472**
			bsky_phi	0.375	2.429	0.205
		gemma_27	ant	0.448	2.274	0.249
			bsky	**0.363**	**1.826**	**0.538**
			bsky_phi	0.438	2.231	0.385
		llama8	ant	0.528	2.096	0.413
			bsky	0.489	2.306	0.256
			bsky_phi	**0.406**	**1.918**	**0.482**

Table 12: Aggregated Performance by Demographic and Response Surveyor Models

Wave	Demo Surveyor	Resp. Surveyor	Backstory	EMD ↓	Frob. ↓	Cron. α ↑
All	gemma_12	gemma_12	ant	0.36	**2.22**	**0.38**
			bsky	**0.35**	2.23	0.36
			bsky_phi	0.42	2.50	0.24
		gemma_27	ant	0.38	2.46	0.26
			bsky	0.37	**2.10**	**0.46**
			bsky_phi	**0.34**	2.27	0.36
		llama8	ant	0.41	**2.29**	0.34
			bsky	**0.40**	2.36	0.31
			bsky_phi	0.43	2.30	**0.34**
All	gemma_27	gemma_12	ant	0.37	2.33	0.31
			bsky	**0.32**	**2.19**	**0.39**
			bsky_phi	0.40	2.27	0.36
		gemma_27	ant	0.37	2.30	0.38
			bsky	**0.34**	**2.27**	**0.40**
			bsky_phi	0.35	2.27	0.37
		llama8	ant	0.43	**2.09**	**0.42**
			bsky	**0.40**	2.33	0.32
			bsky_phi	0.42	2.34	0.31
All	llama8	gemma_12	ant	0.36	2.42	0.31
			bsky	**0.32**	**2.04**	**0.47**
			bsky_phi	0.36	2.46	0.26
		gemma_27	ant	0.41	2.33	0.30
			bsky	**0.35**	**1.93**	**0.50**
			bsky_phi	0.39	2.26	0.43
		llama8	ant	0.44	2.24	0.39
			bsky	0.44	2.44	0.29
			bsky_phi	**0.42**	**2.21**	**0.44**

Wave	Exp.	EMD (5-pt / 3-pt) ↓	Frob. (5-pt / 3-pt) ↓	Cron. α (5-pt / 3-pt) ↑
34	$\text{SYN}_{\text{Gemma}}$	**0.3506 / 0.1899**	**2.3012 / 1.8723**	**0.3917 / 0.3666**
34	$\text{ANT}_{\text{LLaMa}}$	0.3543 / 0.2049	2.4578 / 2.0427	0.3351 / 0.2957
34	SYN_{Phi}	0.3805 / 0.2110	2.4333 / 1.9997	0.3785 / 0.3449
99	$\text{SYN}_{\text{Gemma}}$	**0.3824 / 0.2240**	**2.0842** / 1.8628	**0.3881 / 0.3161**
99	$\text{ANT}_{\text{LLaMa}}$	0.4369 / 0.2506	2.1485 / **1.8518**	0.3443 / 0.3107
99	SYN_{Phi}	0.4079 / 0.2286	2.2112 / 1.9212	0.3115 / 0.2686

Table 13: Sensitivity analysis under a coarser 3-point ordinal response scale. Across the six primary wave-level comparisons (three metrics $\times$ two waves), $\text{SYN}_{\text{Gemma}}$ remains the strongest overall system in five cases, indicating that the main findings are robust to response-scale resolution.

Metric	Mean	Median	Min	Max
Pearson r	0.8937	0.9136	0.7045	0.9820

Table 14: Correlation between the 5-point and 3-point evaluations across settings. In addition to the high Pearson correlation shown above, we observe strong rank preservation with Spearman $\rho \approx 0.87$.