THE NEED FOR SOCIALLY-GROUNDED PERSONAS FOR USER SIMULATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Synthetic personas are widely used to condition large language models (LLMs) for social simulation, yet most personas are still constructed from coarse sociodemographic attributes or summaries. We revisit persona creation by introducing **SCOPE**, a socially grounded framework for persona construction and evaluation, built from a 141-item, two-hour sociopsychological protocol collected from 124 U.S.-based participants. Across seven models, we find that demographic-only personas are a structural bottleneck: demographics explain only $\sim$1.5% of variance in human response similarity. Adding sociopsychological facets improves behavioral prediction and reduces over-accentuation, and non-demographic personas based on traits and identity achieve strong alignment with substantially lower bias. These trends generalize to SimBench (441 aligned questions), where SCOPE personas outperform default prompting and NVIDIA Nemotron personas, and SCOPE augmentation improves Nemotron-based personas. Our results indicate that persona quality depends on sociopsychological structure rather than demographic templates or summaries.

1 INTRODUCTION

LLMs and agentic AI systems are increasingly deployed in settings where they must reason about, simulate, or stand in for human behavior Batzner et al. (2025); Gui & Toubia (2023). These include conversational assistants, recommender systems, safety and fairness evaluation, policy analysis, and agent-based social simulation Mullick et al. (2024); Benary et al. (2023); Qiu et al. (2025). To execute these tasks, recent work has adopted **synthetic personas**: *structured representations of hypothetical individuals used to condition model outputs toward particular traits, identities, or behavioral tendencies* Zhang et al. (2018); Park et al. (2023). Despite their growing importance, the construction of personas in NLP and AI has remained narrow and under-theorized Gupta et al. (2024). As summarized in Table 4, most existing approaches rely on short free-text descriptions, small sets of sociodemographic attributes, or model-written summaries Batzner et al. (2025); Venkit et al. (2025b). Even large-scale persona resources such as PersonaHub Ge et al. (2024), Nemotron Meyer & Corneil (2025), and related systems treat demographics or brief identity cues as proxies for human individuality Kane & Schubert (2023). Table 1 further shows that these frameworks capture only limited slices of sociopsychological structure, typically omitting psychological traits, values, and behavioral patterns that social science research identifies as central drivers of behavior Turner & Reynolds (2012); Costa Jr & McCrae (2000). This reliance on demographic-centric abstractions persists despite evidence that demographics alone are weak predictors of individual-level behavior Salminen et al. (2019); Khan et al. (2024).

We therefore revisit persona creation to address both gaps by introducing **SCOPE** (**S**ociopsychological **C**onstruct of **P**ersona **E**valuation), a human-grounded framework for constructing and evaluating synthetic personas. Motivated by the limitations of existing paradigms (Table 4), SCOPE models personas as multidimensional sociopsychological profiles rather than demographic templates or narrative summaries. We collect rich persona data from 124 U.S.-based participants using a two-hour, 141-item protocol spanning **eight facets**: *demographics, sociodemographic behavior, values and motivations, personality traits, behavioral patterns, identity narratives, professional identity*, and *creativity*. Importantly, SCOPE also introduces an evaluation paradigm centered on *structural fidelity*. Instead of asking whether a persona reproduces individual answers verbatim, we ask whether it reproduces the *pattern* of human responses. We perform this using correlation-based

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

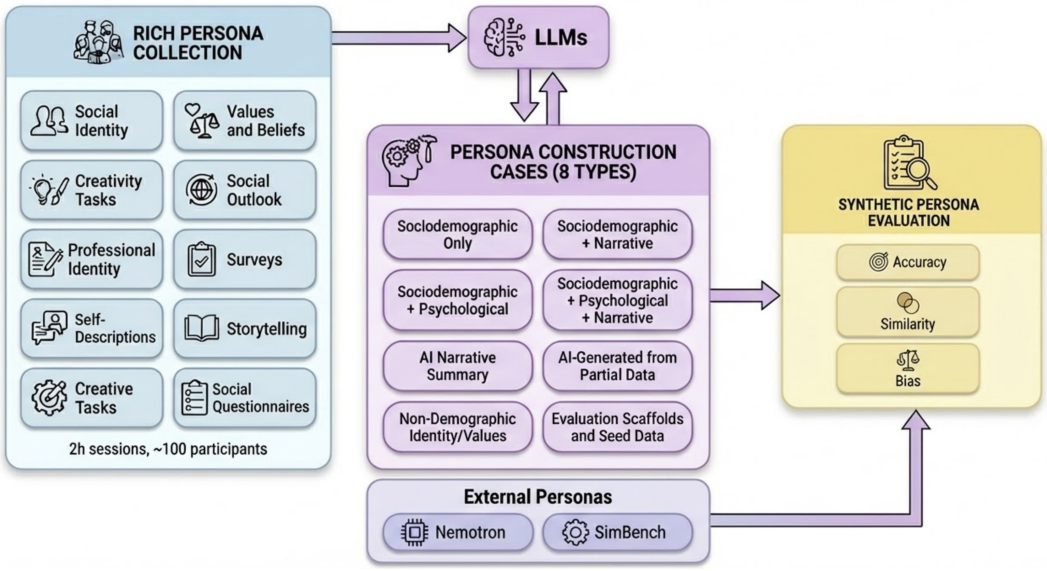


Figure 1: **End-to-end pipeline of our persona construction and evaluation framework.** The three major stages are: (1) **Rich Persona Collection**: a two-hour, 141-item sociopsychological survey capturing eight facets across 124 participants; (2) **Persona Construction Cases**: structured persona representations generated by selectively combining demographic, behavioral, psychological, and narrative facets, as well as AI-generated summaries; (3) **Synthetic Persona Evaluation**: measuring accuracy, similarity to human responses, and group bias.

alignment metrics, complemented by exact-match accuracy and a formal measure of demographic bias accentuation.

Across seven model families[1] and a broad set of persona construction strategies, our results reveal a consistent pattern. Demographic-only personas are a structural bottleneck: demographic similarity explains only $\sim$1.5% of behavioral variance among humans, yet demographic-only prompting more than doubles this signal, over-generalizing group structure. Adding sociopsychological grounding steadily improves alignment while reducing this over-accentuation, and non-demographic personas built from traits and identity can match full conditioning while substantially reducing bias. Longer model-written summaries do not substitute for structured, human-grounded facets: persona *structure* matters as much as information quantity. We further validate these findings on SimBench Hu et al. (2026), a large external benchmark of social and behavioral questions, where SCOPE personas outperform default prompting and NVIDIA Nemotron personas. Most importantly, SCOPE works as a modular augmentation layer: adding SCOPE facets to off-the-shelf Nemotron personas raises their SimBench accuracy from 0.496 to 0.533, above SimBench's default personas (0.436). Unlike prior grounded-persona methods that each propose a new way to build a persona (Lee et al., 2025; Joshi et al., 2025; Moon et al., 2024; Park et al., 2024), SCOPE is a diagnostic and augmentation framework: it measures when a persona over-accentuates demographics and enriches the persona pipelines practitioners already use Venkit et al. (2025a). Throughout, our claims concern aggregate response structure across people, not the exact prediction of any single individual.

2 Related Work

2.1 What Constitutes a Persona?

In research fields of sociology and psychology, a persona is not reducible to a demographic profile or a short narrative description. Instead, human behavior is understood as emerging from interacting layers of social, psychological, and contextual structure. Social Identity Theory emphasizes that individuals

[1]GPT4o, GPT5.1, Claude 3.5-Sonnet, Gemini 2.0 Flash, Gemini 2.5 Pro Thinking, DeepSeek R1, Qwen3

Persona Data Type / Facet	Persona Hub	Nemotron	SPeCtrum	Stanford 1000	SCOPE (Ours)
Demographic Information	●	●	●	●	●
Sociodemographic Behavior	×	×	○	○	●
Values & Motivations	×	×	×	○	●
Personality Traits (Big Five)	×	×	×	×	●
Behavioral Patterns & Preferences	×	×	○	○	●
Identity Narratives	●	○	●	●	●
Professional Identity	×	●	×	○	●
Creativity & Cognitive Flexibility	×	×	×	×	●
Grounding in Human Data	×	×	×	●	●
Sociopsychological Structure	×	×	×	×	●
Designed for Evaluation	×	●	×	×	●

● = *information present;* × = *information absent;* ○ = *partially or implicitly supported.*

Table 1: Comparison of major persona construction approaches (Persona Hub Ge et al. (2024), Nemotron Meyer & Corneil (2025), SPeCtrum Lee et al. (2025), and Stanford 1000 Park et al. (2024)). SCOPE is the only framework that integrates all eight sociopsychological facets and explicitly separates conditioning versus evaluation dimensions.

derive meaning and norms from group memberships (e.g., race, gender, profession), but does not treat demographic categories as deterministic predictors of individual behavior (Tajfel & Turner, 2001; Turner & Reynolds, 2012). Personality psychology further demonstrates that stable traits, most prominently the Big Five, predict consistent patterns of cognition, affect, and behavior across contexts (Costa Jr & McCrae, 2000; Costa & McCrae, 1992). Complementing traits, value theories such as Schwartz's universal values framework show that motivational priorities systematically shape moral judgments, preferences, and life choices as well (Schwartz, 1992). Beyond traits and values, narrative identity theory shows that individuals organize their experiences through internalized life stories, which provide meaning, and a goal structure (McAdams, 1995; 2001). These narratives influence how people interpret situations and decisions, particularly in ambiguous or moral domains. Finally, person–situation interaction models emphasize that behavior is best understood as structured patterns across situations rather than single responses in isolation (Mischel & Shoda, 1995).

2.2 Persona Construction in NLP

Recent work constructs personas as short textual profiles or self-descriptive sentences appended to dialogue, popularized by PersonaChat-style datasets and models (Zhang et al., 2018). More recent persona resources extend this approach by generating personas from sociodemographic attributes, often aligned with census statistics. Examples include Persona Hub (Ge et al., 2024) and NVIDIA's Nemotron Persona Meyer & Corneil (2025), which provides millions of fully synthetic U.S. personas. While effective for scale, these approaches largely treat personas as demographic or lightly behavioral templates. Empirical analyses increasingly show that such representations risk collapsing behavioral diversity into demographic stereotypes Venkit et al. (2025b); Cheng et al. (2023); Guo et al. (2025). A growing body of work also argues that persona quality depends on psychologically meaningful structure. SPeCtrum models identity as a composition of social identity, personal identity, and life context (Lee et al., 2025). PB&J incorporates psychologically grounded rationales to improve preference and judgment prediction (Joshi et al., 2025). ValueSim focuses on simulating individual value systems via structured backstories (Du et al., 2025). However, with an increasingly diverse approach in persona creation, there is less clarity on what facets or approaches actually represent users.

Synthetic personas are also increasingly used as virtual participants in social science, policy analysis, and system evaluation. SimBench aggregates large-scale social and behavioral datasets to benchmark whether LLMs can reproduce population-level response patterns (Hu et al., 2026). Other work evaluates LLMs as substitutes for human survey respondents, highlighting their promise and limitations (Argyle et al., 2023; Kolluri et al., 2025). In recommender systems, persona-based modeling has been shown to improve preference prediction when grounded in psychologically meaningful features rather than demographics alone (Joshi et al., 2025; Ketipov et al., 2023).

3 SOCIALLY GROUNDED PERSONA FRAMEWORK

Synthetic personas are widely used in NLP and AI, yet their construction remains under-theorized and weakly grounded in empirical social science. Prior work in computational personality and behavior modeling shows that psychological traits and values are substantially more predictive of behavior than sociodemographic variables alone, a distinction largely absent from current persona construction practices Ketipov et al. (2023); Joshi et al. (2025). To address this gap, we introduce **SCOPE: Sociopsychological Construct of Persona Evaluation**, a sociopsychologically informed framework for constructing rich, multidimensional personas.

3.1 THE SCOPE FRAMEWORK

To move beyond demographic-only personas, SCOPE introduces a multidimensional sociopsychological architecture grounded in established findings from social psychology, personality science, narrative identity theory, and computational behavior modeling. SCOPE defines a structured persona representation consisting of **eight facets** (Table 10). Each facet is defined through structured questions and free-text inputs. SCOPE distinguishes between two categories of facets:

Conditioning Facets (used for persona construction). These include *Demographics*, *Sociodemographic Behavior*, *Personality Traits*, and *Identity Narratives*, which provide the social, psychological, and narrative grounding for persona generation.

Evaluation Facets (held out during construction). These include *Values & Motivations*, *Behavioral Patterns*, *Professional Identity*, and *Creativity & Innovation*, which serve as behavioral and attitudinal targets for evaluating persona coherence.

Each facet is grounded in established sociological and psychological frameworks. Table 10 details the content for each facet. This partitioning reflects a core principle of sociopsychological modeling: identity is partly observable (e.g., demographics and personality traits) and partly inferential (e.g., values, motivations, and creative expression). Accordingly, SCOPE provides both (1) a mechanism for constructing multidimensional LLM personas (2) a framework for evaluating their behavioral grounding. In total, SCOPE comprises **141 curated attributes** for all facets (Appendix A.7).

3.2 DATA COLLECTION DESIGN

To establish a socially grounded foundation for SCOPE, we designed a two-hour survey capturing detailed sociopsychological data across the eight facets in Table 10. This enables direct comparison with real human responses, unlike prior work that relies primarily on fully synthetic data. The final instrument contains 141 items administered via Typeform[2] and spans nominal, ordinal, and free-text formats. Participants completed *51 Likert-scale items, 37 single-choice dropdown questions, 20 multiple-choice questions, 3 short free-text inputs,* and *30 narrative or creative prompts* (e.g., life stories, identity reflections, and hypothetical reasoning tasks) (Appendix A.6-A.7).

Each facet was informed by established social-science frameworks. Personality items followed the BFI-2-S, value items drew on Schwartz's value theory, and behavioral items were aligned with large-scale surveys such as Pew Internet & Technology, the General Social Survey (GSS), and the World Values Survey (WVS) Davis & Smith (1999); Pew Research Center (2024); Inglehart et al. (2014). All items were written as an original composite survey rather than directly reproducing any single instrument. We recruited a U.S.-based sample guided by Census-like proportions, with explicit emphasis on racial diversity. Recruitment occurred in two stages: (1) outreach via Twitter/X, LinkedIn, and professional Slack groups; and (2) stratified sampling through a commercial research platform (User Interviews). The study protocol was approved by the institution's Ethics Office.

To ensure authentic responses, we employed instructional deterrents and automated screening Zhang et al. (2025); Christoforou et al. (2024). GPTZero was used as a coarse AI-authorship filter, supplemented by manual review Brown & Jensen (2023); gpt (2025). Participants were compensated $50; 124/146 completed the survey with verified human-authored responses, yielding 17,484 total responses. The final sample comprised 33.0% White (41), 19.3% African American or Black (24), 20.1% Hispanic or Latino (25), 14.5% Asian (18), 7.2% American Indian or Alaska Native (9),

[2]https://www.typeform.com/

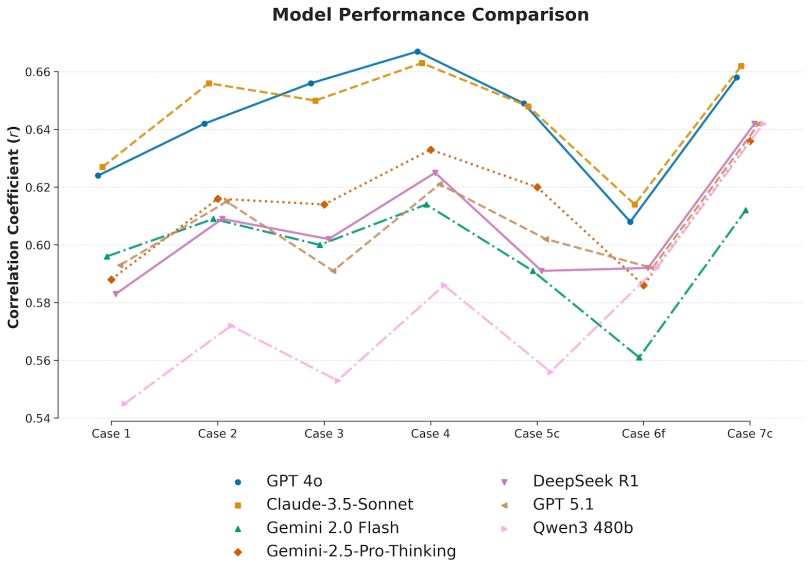


Figure 2: Mean behavioral correlation ($\bar{r}$) with human responses across persona construction cases, shown for all seven models. Richer, multi-facet cases (2–4, 7c) consistently improve over the demographic-only case (1).

2.4% Native Hawaiian or Pacific Islander (3), and 2.2% Other (4). This closely aligns with U.S. Census distributions while modestly oversampling underrepresented groups. Dataset statistics and question-type distributions are in Appendix A.7.

3.3 Synthetic Persona Construction

Given the human-grounded SCOPE profiles, we next construct synthetic personas that can be created in an LLM and systematically varied in representational richness. For each of the 124 participants, we serialize their SCOPE profile into a structured "persona scaffold" using only the conditioning facets (Demographic Information, Sociodemographic Behavior, Personality Traits, Identity Narratives, Table 10), while holding out the remaining facets as evaluation targets. We define facet-ablation variants that differ only in which parts of the human profile are exposed to the model:

Case 0 (No Personas) is a baseline with no persona information, isolating the gain from using personas. **Case 1 (Demography-Only)** uses basic sociodemographic attributes, resembling most current work. **Case 2 (Demography + Narratives)** adds personal and professional identity narratives; **Case 3 (Demography + Traits)** adds Big Five traits from the BFI-2-S; and **Case 4 (Full Conditioning)** uses all conditioning facets (demographics, traits, narratives, and sociodemographic behavior).

We also introduce three **LLM-centric** variants that mirror current automated persona generation. **Case 5 (LLM Narrative Summaries)** gives the model a condensed persona summary it generates itself from the full SCOPE profile, as in Nemotron. **Case 6 (LLM Facet Completion)** provides only partial information (e.g., demographics, or demographics + identity) and asks the model to infer the missing facets; different ablations of the input define the Case 6 subsets. **Case 7 (Non-Demographic Personas)** omits demographics entirely (e.g., narrative-only or trait + narrative), testing whether non-demographic cues alone support behavioral simulation.

Each variant is paired with the same set of held-out SCOPE questions and external task. We show examples of each case in Table 14 in the Appendix.

4 Evaluation Framework

Our evaluation addresses a central question in persona modeling: rather than asking whether model responses *match* human answers, we ask whether synthetic personas exhibit the *same behavioral*

structure. Because humans rarely give identical answers but do show stable patterns across related questions Kambhatla et al. (2022); Venkit et al. (2025b), we emphasize structure-preserving similarity: when a human answers relatively high on some questions and low on others, does the synthetic persona vary in parallel? We evaluate personas along three axes:

A. Behavioral correlation: pattern similarity between human and model responses.
B. Accuracy: exact agreement on social-attribute labels and Likert responses.
C. Bias: extent to which a persona over- or under-accentuates demographic differences relative to real group responses. We look at this in Section 4.3

We first analyze these metrics on our SCOPE persona ablations (Cases 1–7), then evaluate generalization on SimBench Hu et al. (2026), and finally study demographic bias.

4.1 Baseline Persona Comparisons

For each participant i, persona case c, and model m, let $\mathbf{y}_i^{\text{human}} \in \mathbb{R}^K$ denote that participant's vector of held-out SCOPE answers (across K evaluation questions), and $\mathbf{y}_{m,c,i}^{\text{model}} \in \mathbb{R}^K$ the answers generated by model m when instantiated with case-c persona information. We construct a gold-standard set of 72 MCQ items drawn from *sociodemographic behaviours*, *personal values and motivations*, and *behavioral patterns*, used as the first baseline of evaluation.

Correlation. We measure behavioral alignment using the Pearson correlation coefficient

$$r_{m,c,i} = \text{corr}\big(\mathbf{y}_i^{\text{human}},\ \mathbf{y}_{m,c,i}^{\text{model}}\big), \tag{1}$$

and report the participant-level mean

$$\bar{r}_{m,c} = \frac{1}{N}\sum_{i=1}^{N} r_{m,c,i}. \tag{2}$$

Intuitively, $\bar{r}_{m,c}$ answers: *does this persona reproduce the similar response characteristics as its human counterpart, up to a linear transformation?*

Accuracy. For completeness, we also compute an exact-match accuracy over all evaluation items:

$$\text{Acc}_{m,c} = \frac{1}{NK}\sum_{i=1}^{N}\sum_{k=1}^{K} \mathbf{1}\left[y_{m,c,i,k}^{\text{model}} = y_{i,k}^{\text{human}}\right], \tag{3}$$

where equality is defined at the level of the discrete response option (Likert rating, multiple-choice option, or categorical label).

Do richer personas help? Figure 2 shows correlation across all models, and Table 2 summarizes GPT-4o performance[3] across the most informative persona cases; the full seven-model breakdown appears in Tables 15–16. The following trends emerge across models:

I. Demographics alone are insufficient. Demography-only personas (Case 1) achieve only moderate correlation ($\bar{r} = 0.624$) and the lowest accuracy (35.1%), despite being the dominant paradigm in current work.

II. Adding sociopsychological facets improves alignment. Sequentially enriching personas with identity narratives (Case 2), traits (Case 3), and full SCOPE conditioning (Case 4) steadily increases correlation and accuracy, with Case 4 reaching $\bar{r} = 0.667$ and 39.7% accuracy.

III. Human-grounded facets outperform AI-compressed personas. AI-generated narrative summaries (Case 5c) slightly improve over demographics but never surpass fully grounded representations, even when summaries are long. Non-demographic SCOPE personas built only from traits and identity narratives (Case 7c) match or exceed the best full-information case in correlation while avoiding direct use of demographic attributes.

IV. Non-demographic information drives performance. Sociodemographic attributes are not the most important element of a persona: in Case 6 variants, personas built from traits or identity narratives (or both) tend to achieve higher accuracy than those built from demographics alone.

[3]We report GPT-4o as the best-performing model; complete results are in the Appendix.

GPT-4o Persona Case	$\bar{r}$	Acc.	Bias %
0: No Persona	0.448	20.47	-
1: Demography only	0.624	35.07	101.23
2: + Narratives	0.642	37.74	8.30
3: + Traits	0.656	38.60	75.50
4: Full conditioning	0.667	39.67	-6.40
5c: LLM summary (1500w)	0.649	38.07	-38.42
6f: AI identity + traits	0.608	43.42	-40.40
7c: traits + narrative only	0.658	39.68	-56.35

Table 2: Behavioral alignment of GPT-4o personas across key SCOPE cases. $\bar{r}$ is mean Pearson correlation with human responses; Acc. is exact-match accuracy (%). Bias% is defined in Eq. (7). Full results for all models(1–7a/b/c, 5a–d, 6a–f) are in the appendix.

V. Socially grounded personas can simulate sociodemographic and behavioral patterns. Appendix Table 7 shows that, across tasks, personas replicate sociodemographic behaviors and personality traits well, while fine-grained day-to-day behaviors are harder to capture. Personas are thus more reliable for group-level patterns than for individual-level preferences.

Across models the ordering is consistent: richer, multi-facet personas yield higher correlations than demographic-only cases. The gain at each step is small, but it holds across all seven models (Figure 2), so we read the trend, not any single gap, as the result. We additionally explore open-ended questions in Appendix B.4, using the frameworks of Venkit et al. (2025b) and Chakrabarty et al. (2024) to evaluate writing style and diversity of text-based answers. Writing style is highly facet-dependent and sensitive to the conditioning case.

4.2 Task Based Performance Analysis

To test whether our findings generalize beyond our internal questionnaire, we evaluate SCOPE personas on SimBench, a composite benchmark built from 20 large-scale social and behavioral datasets (e.g., AfroBarometer, ESS, ISSP, LatinoBarometro, OpinionQA). We align SimBench items with our personas, retaining 441 questions whose wording, response scale, and sociotechnical context are compatible with the sociodemographic groups in our data.

We created GPT-4o personas for each participant and persona case, then prompted the model to answer the **441 SimBench questions** in character. We report **majority-vote accuracy**: whether the persona selects the most frequent human response for its matching demographic group (within 3 turns). We also compare against personas sampled from the NVIDIA Nemotron USA dataset Meyer & Corneil (2025), using their full textual descriptions and the same GPT-4o backend (examples in Appendix A.10).

SCOPE personas with sociopsychological facets (Cases 2–4 and 7c) outperform Nemotron across the 441 items, while demographic-only SCOPE personas are comparable to Nemotron. This supports our central claim that *one-dimensional sociodemographic personas are insufficient even on external benchmarks*. Augmenting Nemotron with SCOPE facets yields its best variant, showing that *SCOPE can enrich existing personas for more human-like predictions*.

4.3 Bias and Behavioral Disparities

Finally, we assess whether persona prompting *amplifies* or *attenuates* demographic structure relative to human responses. Our aim is not to remove legitimate demographic signal, but to detect when personas *over-accentuate* demographic differences.

Human baseline: Using $N_{\text{human}} = 124$ participants, 13 demographic variables, and 134 evaluation questions, we measure how strongly demographics predict behavior. For each participant pair (i, j) we compute demographic and response similarity via cosine similarity over standardized feature vectors, then take the Pearson correlation between them over all unique pairs.

Persona Source	Case Description	Accuracy
SCOPE (Ours)	4. Full conditioning	**0.584**
	2. Demo. + Narratives	0.557
	6a. AI profile	0.553
	3. Demo. + Traits	0.531
	1. Demographics only	0.518
	5d. AI summary	0.494
Nemotron	**SCOPE augmented**	0.533
	Persona text only	0.496
	All Data	0.446
SimBench	Default personas	0.436
Baseline	No Persona	0.258

Table 3: Task-based evaluation on **SimBench**.We report accuracy using GPT-4o personas under different construction strategies.

Let $\mathbf{x}_i^{\text{demo}}$ and $\mathbf{x}_i^{\text{resp}}$ denote the standardized demographic and response vectors for participant i. We define pairwise similarities and the human baseline as:

$$d_{ij} = \cos\Big(\mathbf{x}_i^{\text{demo}}, \mathbf{x}_j^{\text{demo}}\Big), \quad s_{ij} = \cos\big(\mathbf{x}_i^{\text{resp}}, \mathbf{x}_j^{\text{resp}}\big), \quad r^{\text{human}} = \text{corr}(\{d_{ij}\}_{i<j}, \{s_{ij}\}_{i<j}), \tag{4}$$

flattening the upper triangle of the similarity matrices.

Empirically, we obtain

$$r^{\text{human}} = 0.132 \quad (p < 0.001),$$

implying that demographic similarity explains only $r^2 \approx 1.5\%$ of the variance in response similarity. This baseline is small but sits in the range that psychology treats as substantively meaningful for real-world behavior (Funder & Ozer, 2019), and it sets a conservative reference for how much demographic structure is present in *real* responses.

Demographic accentuation in personas: For each model m and persona construction case c, we compute an analogous coefficient by replacing human response similarity with similarity between the corresponding AI persona answers:

$$r_{m,c}^{\text{AI}} = \text{corr}\big(d_{ij},\, s_{m,c,ij}^{\text{AI}}\big). \tag{5}$$

Demographic accentuation is the difference from the human baseline and its normalized percentage:

$$\Delta r_{m,c} = r_{m,c}^{\text{AI}} - r^{\text{human}}, \tag{6}$$

$$\text{Bias}\%_{m,c} = 100 \times \frac{\Delta r_{m,c}}{r^{\text{human}}}. \tag{7}$$

Positive values show persona makes demographically similar individuals *too similar* (over-accentuation), while negative values indicate that the persona *underplays* demographic structure. Because the human baseline is genuinely small (r^{human}=0.132), we report the absolute change Δr alongside the percentage and read them together with the raw correlations. Our *bias* metric captures demographic accentuation relative to human baselines, not normative unfairness or discrimination.

Results across cases and models: Table 2 and Appendix Tables 15–16 show a consistent pattern. Demographic-only prompting strongly over-accentuates demographic similarity: for GPT-4o, Case 1 yields $\text{Bias}\% = 101.23$ (more than doubling the human signal), and Claude-3.5-Sonnet reaches 115.67, indicating that models collapse behavior toward demographic stereotypes. Adding sociopsychological grounding reduces this, though the magnitude varies by facet: for GPT-4o, Case 2 (demographics + identity) falls to near-baseline (8.30), while Case 3 (demographics + traits) still over-accentuates (75.50) because demographics remain explicit inputs, and full conditioning (Case 4) slightly undershoots the baseline (-6.40). Non-demographic personas are the most conservative: Case 7c (traits + identity, no demographics) yields $\text{Bias}\% = -56.35$ while keeping high correlation with human response structure ($\bar{r} = 0.658$). The same ordering holds across the other models.

4.4 GROUP-BASED SCOPE PERSONA ANALYSIS

The SCOPE framework also supports demographic-conditioned evaluation: we can test whether a construction strategy over-depends on a specific social group. We focus on race, which shows the highest demographic correlation (Table 13), and compare AI demographic response correlations against race-conditioned human baselines (Appendix Table 6). This analysis is exploratory (Wagenmakers et al., 2012): some groups are small, so we use it to show what the framework can diagnose rather than to make firm claims about any group.

Human baselines differ by race. The within-group human baseline is moderate for White (r=0.155, n=41) and African American or Black (r=0.208, n=24) participants, weak for Hispanic or Latino participants (r=0.076, n=25), and near zero for Asian (r=−0.004, n=18) and American Indian or Alaska Native (r=−0.005, n=9) participants. How much demographics explain behavior therefore varies widely across groups and is not a reliable predictor on its own. The pattern is itself telling: where a group carries little stable demographic signal, there is little for a persona to ground on, yet demographic-only prompting still manufactures group structure (below).

Bias is case- and race-dependent. Appendix Tables 5–6 report bias accentuation averaged across racial groups and the per-group breakdown. Demographics-only personas (Case 1) produce the largest positive bias, strongest for groups with weak human baselines (Asian and American Indian or Alaska Native), where a persona risks flattening or over-accentuating a group Cheng et al. (2023). Adding narratives (Case 2) or traits (Case 3) reduces bias, and full conditioning (Case 4) brings AI behavior close to the human baseline. Personas that drop demographics entirely yield negative bias: they reduce demographic dependence but can under-represent a group. We analyze race here, but the same conditioning applies to every demographic attribute.

5 DISCUSSION AND CONCLUSION

Demographics are a bottleneck; sociopsychological grounding helps. Personas built solely from demographics (Case 1) consistently underperform across models and on SimBench (Tables 2, 15, 16): in human data, demographic similarity explains only ∼1.5% of response variance, yet demographic-only personas more than double this signal. Adding values, traits, identity narratives, and behavioral signals steadily improves alignment, with full conditioning (Case 4) reaching the highest correlation while reducing bias. These gains are not driven by verbosity: AI-written summaries (Case 5) are longer yet weaker than structured profiles, and non-demographic personas (Case 7c) match or exceed full conditioning while substantially reducing bias. Bias reduction is thus best achieved by replacing coarse demographic proxies with psychologically meaningful structure, not by removing it.

SCOPE as an augmentation framework. Rather than replacing existing persona systems, SCOPE is a modular augmentation layer: adding SCOPE facets to Nemotron personas raises their SimBench accuracy from 0.496 to 0.533, above SimBench's default personas (0.436). To limit re-identification risk, we release only the SCOPE personas, the augmentation and evaluation pipeline, and aggregate statistics in an anonymized repository, from which users can build their own SCOPE-augmented personas. We also release one million synthetic SCOPE personas that outperform existing frameworks on behavioral and social tasks.

LIMITATIONS

We note four limitations. First, the participant pool is U.S.-based, so culturally contingent norms may not transfer globally despite Census-like sampling. Second, our fixed survey instrument cannot cover all drivers of behavior, such as longitudinal or situational context. Third, correlation-based evaluation captures relative response structure but does not guarantee causal fidelity or calibration.

AI USE STATEMENT

We used generative AI tools only for grammatical and stylistic editing of the manuscript. We did not use generative AI tools for research ideation, experimental design, analysis, or the interpretation

of results, these are our own. LLMs also appear as the object of study in our experiments (persona generation and simulation), and their use in that setting is described in full in Sections 3–4. We reviewed all AI-assisted text.

Ethics Statement

We collected a two-hour sociopsychological questionnaire from U.S.-based adult participants; the protocol underwent internal ethical review before deployment. Participants gave informed consent and were compensated $50. To reduce privacy risk, we did not collect direct identifiers (legal names, email addresses, phone numbers, street addresses), and we did not collect protected health or financial identifiers. Free-text responses were stored under anonymized participant IDs and screened for accidental disclosure of sensitive information. The dataset used for modeling contains only aggregated demographic categories and de-identified text, and all analyses were run on this de-identified corpus. Because persona conditioning can amplify stereotypes, we explicitly measure demographic accentuation (Section 4.3) and highlight persona variants that maintain behavioral alignment while reducing demographic clustering. This work does not claim that models can replace or faithfully reproduce individual humans: our evaluation targets aggregate behavioral patterns, not individual-level substitution. We position SCOPE as guidance for building more realistic and more responsible personas.

Reproducibility Statement

The SCOPE framework, facet definitions, and question-type composition are described in Section A (Table 10) and Appendices A.6–A.7. The persona construction cases and their exact information content are defined in Section 3.3 and illustrated in Table 14. All evaluation metrics (correlation, accuracy, and bias accentuation) are defined with equations in Sections 4 and 4.3, and full per-model results are reported in Appendix Tables 15–16. The SimBench alignment procedure and external data formats are documented in Section 4.2 and Appendix A.10. We release the persona data, construction and evaluation code, and augmentation pipeline in an anonymized repository (see the note in Section 5).

References

Gptzero: Ai detection tool. `https://gptzero.me/`, 2025. Accessed 2025-12-30.

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.

Jan Batzner, Volker Stocker, Bingjun Tang, Anusha Natarajan, Qinhao Chen, Stefan Schmid, and Gjergji Kasneci. Whose personae? synthetic persona experiments in llm research and pathways to transparency. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pp. 343–354, 2025.

Manuela Benary, Xing David Wang, Max Schmidt, Dominik Soll, Georg Hilfenhaus, Mani Nassir, Christian Sigler, Maren Knödler, Ulrich Keller, Dieter Beule, Ulrich Keilholz, Ulf Leser, and Damian T Rieke. Leveraging large language models for decision support in personalized oncology. *JAMA Network Open*, 6(11):e2343689–e2343689, 2023.

David W Brown and Dean Jensen. Gptzero vs. text tampering: The battle that gptzero wins. In *Proceedings of the International Conference on Social and Education Sciences (IConSES)*, Las Vegas, NV, 2023. International Society for Technology, Education, and Science (ISTES). ERIC Document ED656089.

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. Art or artifice? large language models and the false promise of creativity. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–34, 2024.

Myra Cheng, Esin Durmus, and Dan Jurafsky. Marked personas: Using natural language prompts to measure stereotypes in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1504–1532, 2023.

Evgenia Christoforou, Gianluca Demartini, and Jahna Otterbacher. Generative ai in crowdwork for web and social media research: A survey of workers at three platforms. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pp. 2097–2103, 2024.

Paul T Costa and Robert R McCrae. *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI)*. Psychological Assessment Resources, 1992.

Paul T Costa Jr and Robert R McCrae. NEO personality inventory. In Alan E Kazdin (ed.), *Encyclopedia of Psychology, Vol. 5*, pp. 407–409. American Psychological Association, 2000.

James Allan Davis and Tom William Smith. *General Social Surveys 1972-2000: Cumulative Codebook*. Inter-university Consortium for Political and Social Research, 1999.

Bangde Du, Ziyi Ye, Zhijing Wu, Jankowska Monika, Shuqi Zhu, Qingyao Ai, Yujia Zhou, and Yiqun Liu. Valuesim: Generating backstories to model individual value systems. *arXiv preprint arXiv:2505.23827*, 2025.

David C Funder and Daniel J Ozer. Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, 2(2):156–168, 2019. doi: 10.1177/2515245919847202.

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

George Gui and Olivier Toubia. The challenge of using llms to simulate human behavior: A causal inference perspective. *arXiv preprint arXiv:2312.15524*, 2023.

Hangzhi Guo, Pranav Narayanan Venkit, Eunchae Jang, Mukund Srinath, Wenbo Zhang, Bonam Mingole, Vipul Gupta, Kush R Varshney, S Shyam Sundar, and Amulya Yadav. Exposing ai bias by crowdsourcing: Democratizing critique of large language models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pp. 1169–1180, 2025.

Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. Bias runs deep: Implicit reasoning biases in persona-assigned llms. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

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

Ronald Inglehart, Christian Haerpfer, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Juan Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, and Bi Puranen. *World Values Survey: Round Six - Country-Pooled Datafile*. JD Systems Institute, Madrid, Spain, 2014.

Brihi Joshi, Xiang Ren, Swabha Swayamdipta, Rik Koncel-Kedziorski, and Tim Paek. Improving language model personas via rationalization with psychological scaffolds. *arXiv preprint arXiv:2504.17993*, 2025.

Gauri Kambhatla, Ian Stewart, and Rada Mihalcea. Surfacing racial stereotypes through identity portrayal. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pp. 1604–1615, 2022.

Benjamin Kane and Lenhart Schubert. We are what we repeatedly do: Inducing and deploying habitual schemas in persona-based responses. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10998–11016, 2023.

Rumen Ketipov, Vera Angelova, Lyubka Doukovska, and Roman Schnalle. Predicting user behavior in e-commerce using machine learning. *Cybernetics and Information Technologies*, 23(3):89–101, 2023.

Azal Ahmad Khan, Sayan Alam, Xinran Wang, Ahmad Faraz Khan, Debanga Raj Neog, and Ali Anwar. Mitigating sycophancy in large language models via direct preference optimization. In *2024 IEEE International Conference on Big Data (BigData)*, pp. 1664–1671. IEEE, 2024.

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*, pp. 30084–30099, 2025.

Keyeun Lee, Seo Hyeong Kim, Seolhee Lee, Jinsu Eun, Yena Ko, Hayeon Jeon, Esther Hehsun Kim, Seonghye Cho, Soeun Yang, Eun-mee Kim, Hajin Lim, et al. Spectrum: A grounded framework for multidimensional identity representation in llm-based agent. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6971–6991, 2025.

Dan P McAdams. What do we know when we know a person? *Journal of Personality*, 63(3):365–396, 1995.

Dan P McAdams. The psychology of life stories. *Review of general psychology*, 5(2):100–122, 2001.

Yev Meyer and Dane Corneil. Nemotron-Personas-USA: Synthetic personas aligned to real-world distributions, June 2025. URL `https://huggingface.co/datasets/nvidia/Nemotron-Personas-USA`.

Walter Mischel and Yuichi Shoda. A cognitive-affective system theory of personality: reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological review*, 102(2):246–268, 1995.

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*, pp. 19864–19897, 2024.

Ankan Mullick, Sombit Bose, Rounak Saha, Ayan Bhowmick, Pawan Goyal, Niloy Ganguly, Prasenjit Dey, and Ravi Kokku. On the persona-based summarization of domain-specific documents. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14291–14307, 2024.

Joon Sung Park, Joseph O'Brien, Carrie Jun 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*, pp. 1–22, 2023.

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. *arXiv preprint arXiv:2411.10109*, 2024.

Pew Research Center. Internet & technology. `https://www.pewresearch.org/internet/`, 2024. Accessed 2024.

Haoyi Qiu, Yilun Zhou, Pranav Narayanan Venkit, Kung-Hsiang Huang, Jiaxin Zhang, Nanyun Peng, and Chien-Sheng Wu. Seeing is believing? evaluating vision-language model susceptibility in agent-to-agent multimodal persuasion. *arXiv preprint arXiv:2510.22768*, 2025.

Joni Salminen, Soon-Gyo Jung, and Bernard J Jansen. Detecting demographic bias in automatically generated personas. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–6, 2019.

Shalom H Schwartz. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In *Advances in experimental social psychology*, volume 25, pp. 1–65. Elsevier, 1992.

Henri Tajfel and John C Turner. An integrative theory of intergroup conflict. In Michael A Hogg and Dominic Abrams (eds.), *Intergroup Relations: Essential Readings*, pp. 94–109. Psychology Press, 2001.

Paradigm	Description	Captured Dimensions	Representative Example
Text-Based Descriptions	Short free-text summaries describing personality, background, or preferences.	Low: narrative-only; limited demographic grounding.	Ge et al. (2024)
Sociodemographic Inputs	Personas constructed from attributes such as age, race, gender, etc	Low–medium: demographic-only.	Meyer & Corneil (2025)
Demographic + Identity Signals	Adds identity questions, cultural background, or identity markers on top of demographic variables.	Medium: demographics + self-identity cues.	Lee et al. (2025)
Interview-Based Personas	Personas formed from interview transcripts or long-form questionnaire responses.	Medium–high: narrative + behavioral profiles.	Park et al. (2023)
Our Framework	Personas from sociopsychological data across eight facets.	Facet sociopsychological profiles.	*This work*

Table 4: Overview of major persona construction frameworks in current LLM research. Example of each of the framework is present in Table 12–14 in the Appendix.

John C Turner and Katherine J Reynolds. Self-categorization theory. *Handbook of theories of social psychology*, 2:399–417, 2012.

Pranav Narayanan Venkit, Philippe Laban, Yilun Zhou, Kung-Hsiang Huang, Yixin Mao, and Chien-Sheng Wu. Deeptrace: Auditing deep research ai systems for tracking reliability across citations and evidence. *arXiv preprint arXiv:2509.04499*, 2025a.

Pranav Narayanan Venkit, Jiayi Li, Yingfan Zhou, Sarah Rajtmajer, and Shomir Wilson. A tale of two identities: An ethical audit of human and ai-crafted personas. *arXiv preprint arXiv:2505.07850*, 2025b.

Eric-Jan Wagenmakers, Ruud Wetzels, Denny Borsboom, Han L J van der Maas, and Rogier A Kievit. An agenda for purely confirmatory research. *Perspectives on Psychological Science*, 7(6):632–638, 2012.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 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)*, pp. 2204–2213, 2018.

Simone Zhang, Janet Xu, and AJ Alvero. Generative ai meets open-ended survey responses: Research participant use of ai and homogenization. *Sociological Methods & Research*, 2025. doi: 10.1177/00491241251327130.

A Appendix

This appendix provides additional methodological detail and extended empirical results that support the core claims of the paper but are omitted from the main text for space and clarity. In particular, we present (i) a complete breakdown of the sociopsychological facets underlying SCOPE, (ii) the end-to-end persona construction and evaluation pipeline, and (iii) full model-level results across all persona construction cases.

A.1 Persona construction paradigms

Table 4 gives a broad overview of the major persona construction paradigms in current LLM research, complementing the facet-level comparison in Table 1 of the main text.

A.2 Group-based bias by race

Table 5 reports bias accentuation averaged across racial groups by persona construction strategy, and Table 6 gives the per-group breakdown of AI demographic response correlations by case and race. Both support the group-based analysis in Section 4.4.

Case	Group Bias	Std. Dev.
Case_1: Demographics Only	+0.237	0.043
Case_2: Demographics + Narratives	+0.070	0.113
Case_3: Demographics + Traits	+0.165	0.028
Case_4: Full Information	+0.030	0.059
Case_5a: AI Summary (300–500w)	-0.042	0.084
Case_5b: AI Summary (500–1000w)	-0.033	0.094
Case_5c: AI Summary (1000–1500w)	+0.039	0.139
Case_5d: AI Summary (No Limit)	-0.037	0.121
Case_7a: Narratives Only	-0.049	0.145
Case_7b: Traits Only	-0.068	0.102
Case_7c: Narratives + Traits	-0.053	0.088

Table 5: Average group-based bias (here, Race) accentuation by persona construction strategy across racial groups. Only officially defined persona cases (Cases 1–5 and Case 7) are reported. Positive values indicate increased demographic dependence relative to humans; negative values indicate reduced dependence.

Persona Case	White	African American	Hisp/Lat	Asian	AI / AN
Human baseline	0.155	0.208	0.076	−0.004	−0.005
Dem. (#1)	0.446 (+0.291)	0.449 (+0.242)	0.252 (+0.176)	0.216 (+0.220)	0.250 (+0.255)
Dem.+Narratives (#2)	0.226 (+0.071)	0.183 (−0.025)	0.040 (−0.036)	0.096 (+0.100)	0.237 (+0.241)
Dem.+Traits (#3)	0.341 (+0.186)	0.384 (+0.176)	0.236 (+0.160)	0.115 (+0.119)	0.181 (+0.186)
Full Value (#4)	0.161 (+0.006)	0.175 (−0.033)	0.065 (−0.011)	0.095 (+0.099)	0.081 (+0.086)
AI Summary (#5C)	0.108 (−0.047)	0.137 (−0.071)	0.037 (−0.039)	0.085 (+0.089)	0.257 (+0.262)
Identity Only (#2D)	0.037 (−0.118)	0.057 (−0.151)	−0.064 (−0.140)	−0.038 (−0.034)	0.192 (+0.197)

Table 6: AI demographic response correlations (r) by *persona construction case* and race. The first row reports race-conditioned human baselines. Values in parentheses denote bias accentuation relative to the corresponding human baseline. Positive values indicate demographic over-accentuation; negative values indicate reduced demographic dependence. AI / AN = American Indian or Alaska Native, Hisp/Lat = Hispanic or Latino.

A.3 Facet-level accuracy across cases

Table 7 reports facet-level accuracy scores (sociodemographic behavior, behavioral patterns, and personality traits) across persona construction cases, supporting Finding V in the main text.

A.4 Facet-level grounding and coverage.

Table 10 details the eight sociopsychological facets that constitute the SCOPE framework, including their behavioral motivation, question counts, and theoretical grounding. As discussed in the main paper, SCOPE explicitly distinguishes between facets used for *persona conditioning* (observable or self-reported attributes such as demographics, personality traits, and identity narratives) and facets reserved for *evaluation* (values, behavioral patterns, professional identity, and creativity).

A.5 End-to-end Pipeline.

Figure 1 visualizes the complete SCOPE workflow, from human data collection to persona construction, evaluation, and aggregation. The pipeline highlights how human-grounded sociopsychological data are serialized into multiple persona variants, evaluated against held-out behavioral targets, and summarized through correlation, accuracy, and bias metrics.

A.6 Survey instrument Overview

We used a 141-item sociopsychological questionnaire spanning eight facets: Demographic Information (13), Sociodemographic Behavior (37), Values & Motivations (22), Personality Traits (30), Behavioral Patterns & Preferences (13), Personal Identity & Life Narratives (10), Professional Identity & Career (10), and Creativity & Innovation (6). The instrument mixes discrete-choice and scale-based items with longer narrative prompts to capture both structured attitudes and expressive self-narratives.

Case	Socio.	Behavioral	Traits
Case 1	0.4084	0.2172	0.3356
Case 2	0.4292	0.2206	0.3792
Case 3	0.4299	0.2301	0.3972
Case 4	**0.4438**	**0.2383**	**0.4051**
Case 5a	0.4287	0.2166	0.3858
Case 5b	0.4313	0.2213	0.3866
Case 5c	0.4280	0.2200	0.3897
Case 5d	0.4257	0.2247	0.3957

Table 7: Facet-level accuracy scores across persona construction cases for Socio.: *sociodemographic behavior*, Behavioral: *behavioral patterns*, and Traits: *personality traits*.

Question Type	Frequency	%
Likert-scale (1–5)	51	36.17
Dropdown (single-choice)	37	26.24
Multiple-choice	20	14.18
Open-ended (narrative/creative)	30	21.28
Short free-text	3	2.13
Total	**141**	**100.00**

Table 8: Distribution of question types in the SCOPE questionnaire. Percentages are computed over the full 141-item survey and illustrate the balance between structured (Likert, multiple-choice) and expressive (open-ended) response formats.

For space, we do not reproduce the full questionnaire; instead, Table 11 provides one illustrative example item per facet along with response formats.

A.7 Survey Statistics

Our survey contains 141 questions across eight facets (Table 10). The question-type composition is intentionally mixed to support both structured evaluation and open-ended behavioral analysis: **51 opinion-scale items (1–5 Likert), 37 dropdown-scale items, 30 multi-format narrative prompts, 20 multiple-choice items, and 3 short-text items.**

Facet-wise, the instrument includes 13 demographic items, 37 sociodemographic-behavior items, 22 values/motivations items, 30 Big Five trait items, 13 behavioral-pattern items, 20 identity prompts (10 identity narrative prompts, 10 professional-identity prompts), and 6 creativity prompts. Table 8–9 shows the frequency distribution of the types of questions.

For narrative responses, we report (i) token/word-count distributions by facet (identity, professional, creativity), including median and interquartile range, (ii) missingness rates (items answered as "N/A" or left blank), and (iii) per-participant completion time statistics. These summaries are included to support reproducibility and to contextualize model performance on facets with substantially different response entropy and annotation density.

A.8 Framework Examples

We show how each of the persona framework differs by showcasing examples of each framework in Table 12. We see how our framework contains more details and information to process behavioural and social tasks better.

A.9 Complete model-level results.

Tables 15 and 16 report the full set of experimental results across all seven model families and all persona construction cases. These tables expand on the summarized findings in Section 4, showing that the same qualitative trends, demographic over-accentuation, gains from sociopsychological

SCOPE Facet	Freq	%
Demographic Information	13	9.22
Sociodemographic Behavior	37	26.24
Values & Motivations	22	15.60
Personality Traits (Big Five)	30	21.28
Behavioral Patterns & Preferences	13	9.22
Identity Narratives	10	7.09
Professional Identity	10	7.09
Creativity & Innovation	6	4.26
Total	**141**	**100.00**

Table 9: Distribution of questions across the eight sociopsychological facets in the SCOPE questionnaire. Percentages are computed over the full 141-item instrument.

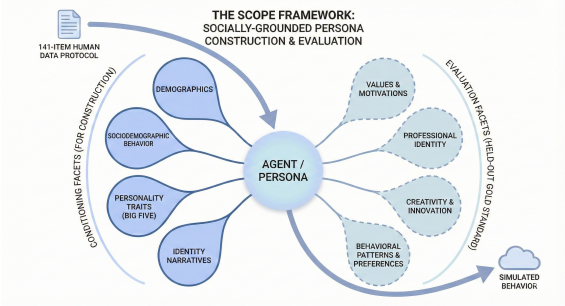


Figure 3: Figure illustrates the facet used for constructing the persona to the ones used for evaluation as a baseline in creating better personas.

grounding, and the robustness of non-demographic personas, hold consistently across models with very different architectures and training regimes.

A.10 External Data Formats: Nemotron Personas and SimBench Instances

To contextualize our SimBench evaluation and Nemotron comparisons, we summarize the *data record formats* used by each external resource.

Nemotron-Personas-USA (persona text + structured attributes). The NVIDIA Nemotron-Personas-USA Meyer & Corneil (2025) dataset[4] consists of fully synthetic U.S. personas generated to mirror census-grounded demographic distributions and provide diverse, non-PII persona descriptions.[5] Each row contains (i) multiple *domain-specific persona paragraphs* (e.g., professional, travel, arts), (ii) a *general persona* summary, and (iii) structured demographic and location attributes (e.g., age, sex, education, occupation, state). Concretely, the dataset viewer exposes fields including: *uuid*, *persona*, *professional_persona*, *sports_persona*, *arts_persona*, *travel_persona*, *culinary_persona*, *cultural_background*, *skills_and_expertise* (and list variant), *hobbies_and_interests* (and list variant), *career_goals_and_ambitions*, along with *sex*, *age*, *marital_status*, *education_level*, *bachelors_field*, *occupation*, *city*, *state*, *zipcode*, *country*.

SimBench (group prompt template + response distribution). SimBench Hu et al. (2026) is a benchmark of *group-level human response distributions* compiled from multiple social and behavioral datasets into a unified format. Each instance provides: (i) a *persona/group prompt template* that describes the respondent group (population-level or demographically grouped), (ii) a *variable map* used to fill template placeholders (empty for population templates), (iii) the *question text*, and (iv) an aggregated *human answer distribution* over discrete options. Specifically, the dataset lists

[4] https://huggingface.co/datasets/nvidia/Nemotron-Personas-USA

[5] The dataset card notes the exclusion of direct identifiers such as names/addresses and describes the generation pipeline via NeMo Data Designer.

Facet	Description	#	Source / References
Demographic Information	Captures foundational sociodemographic attributes (e.g., age, gender, race/ethnicity, education, income, political orientation). Although insufficient to explain behavior on their own, such variables ground broader identity, social context, and structural constraints.	13	*Source:* Social Identity (Demographics) *Refs:* Park et al. (2023)
Sociodemographic Behavior	Measures group-conditioned behavioral tendencies such as media use, digital access, civic engagement, community belonging, and trust in institutions. These behaviors reflect cultural, structural, and lived-experience influences not recoverable from static demographic labels. This facet draws conceptually on Pew Internet & Technology surveys, GSS behavioral modules, and WVS civic participation items.	37	*Source:* Sociodemographic Indicators *Refs:* Park et al. (2023), Pew Research Center (2024), Davis & Smith (1999), Inglehart et al. (2014)
Values & Motivations	Derived from established value frameworks (Schwartz), this facet captures benevolence, tradition, achievement, risk-taking, moral orientation, and long-term goals. Values strongly predict social judgments, moral reasoning, preference formation, and life choices.	22	*Source:* Fundamental Human Values *Refs:* Schwartz (1992), Joshi et al. (2025)
Personality Traits (Big Five)	Encodes stable psychological dispositions, openness, conscientiousness, extraversion, agreeableness, neuroticism, that predict interpersonal behavior, communication style, preference structure, and emotional tendencies. These traits offer robust explanatory power over individual variation.	30	*Source:* Psychological Traits (BFI-2-S) *Refs:* Costa & McCrae (1992), Joshi et al. (2025)
Behavioral Patterns & Preferences	Captures lifestyle habits, emotional responses, motivational strategies, and social tendencies that provide fine-grained behavioral texture. These behaviors emerge from identity, personality, environment, and values, not recoverable from demographic or trait-level information alone.	13	*Source:* Behavioral Patterns *Refs:* Park et al. (2023)
Identity Narratives	Short open-ended reflections on personal history, identity formation, relationships, formative events, professional and self-description. Narrative identity theory suggests that people enact behavior through internalized life stories that shape goals, meaning-making, and social interaction.	10	*Source:* Identity Narrative *Refs:* McAdams (1995), Park et al. (2023)
Professional Identity	Represents occupational roles, work habits, domain knowledge, responsibilities, and career motivations. This facet enables personas to generate domain-relevant reasoning and simulate realistic professional behavior.	10	*Source:* Professional Identity *Refs:* Joshi et al. (2025)
Creativity & Innovation	Assesses expressive flexibility, hypothetical reasoning, narrative originality, and creative decision-making. Creative outputs reflect openness, cognitive flexibility, and a persona's interpretive worldview, providing unique behavioral nuance beyond structured survey items.	6	*Source:* Creativity Tasks *Refs:* Joshi et al. (2025), Cheng et al. (2023), Venkit et al. (2025b)

Table 10: The eight sociopsychological facets of the SCOPE framework. Facets 1–4 form the conditioning inputs used for synthetic persona construction; facets 5–8 serve as held-out evaluation dimensions to assess behavioral grounding, generalization, and fidelity. Question counts derive from the 141-item persona dataset, and facet categories follow the internal persona taxonomy.

primary fields: *dataset_name*, *group_prompt_template*, *group_prompt_variable_map*, *input_template*, *human_answer*, and *group_size*.

Examples: Nemotron records typically provide multiple topical paragraphs that can be concatenated into a single persona prompt (as done in our Nemotron baseline), e.g.:

> **professional_persona:** *[paragraph describing job, work style, responsibilities]*
> **travel_persona:** *[paragraph describing travel preferences and aspirations]*
> **persona:** *[compact overall persona summary]*
> **attributes:** `sex=female, age=28, education_level=high_school, occupation=... , state=WI`

SimBench instances, in contrast, are centered on a group prompt plus a distributional target:

Facet	What it measures	Primary response format(s)	Example item (one per facet)
Demographic Information	Baseline social position and structural context (e.g., age, gender, education, race/ethnicity, political orientation).	Multiple choice; short text.	*"Select Your Age"* (10 bins: *Less than 20 . . . 65 or older*).
Sociodemographic Behavior	Digitally mediated access and civic/media behaviors (internet/AI use, privacy concern, voting importance, institutional trust, cultural belonging).	Dropdown scales (frequency / agreement); some categorical.	*"How often do you use AI tools like ChatGPT or Midjourney?"* (frequency scale: *Several times a day . . . Never*).
Values & Motivations	Motivational priorities and moral/social goals (Schwartz-style value statements; short-/long-term goals).	Likert (1–5) + multi-select goals.	*"It's very important to me to help the people around me. I want to care for others."* (1=*Very Inaccurate . . .* 5=*Very Accurate*).
Personality Traits (Big Five)	Stable dispositions across the Big Five (short-form trait statements; mixture of positively and negatively keyed items).	Likert (1–5).	*"I am organized. I like to keep things in order."* (1=*Very Inaccurate . . .* 5=*Very Accurate*).
Behavioral Patterns & Preferences	Everyday routines and social/goal regulation (motivation under setbacks, recognition response, weekend routines, social comfort).	Mixed: narrative (min word count) + multiple choice.	*"How do you stay motivated when things don't go as planned?"* (open response; 20+ words).
Personal Identity & Life Narratives	Narrative identity: autobiographical self-concept, turning points, relationships, neighborhood context, identity change.	Narrative prompts (min word count).	*Crossroads prompt: "Was there a moment. . . where multiple paths were available. . . ? Tell the whole story."* (open response; 50+ words).
Professional Identity & Career	Occupational self-concept and work ecology (workday structure, success criteria, tools/platforms, time sinks, AI/automation impacts).	Narrative prompts (min word count) + one Likert item.	*"Walk me through a typical workday, highlighting your most critical responsibilities and primary goals."* (open response; 50+ words).
Creativity & Innovation	Expressive flexibility and generative ideation (creative pride, hypothetical projects, invented traditions/holidays, metaphorical self).	Long-form creative prompts (min word count).	*"Invent a holiday and describe how people would celebrate it."* (open response; 100+ words).

Table 11: Questionnaire summary with one illustrative example item per SCOPE facet. The table is intended to communicate the construct coverage and the response modalities used for modeling and evaluation.

group_prompt_template: *"You are a 38 year old Amazon Mechanical Turk worker from the United States."*
input_template: *[question stem + options]*
human_answer: `{A:0.04, B:0.08, C:0.03, D:0.81, E:0.04}`

B Analysis of Open-ended and Writing-style based Questions

In addition to structured multiple-choice and Likert-style evaluation, SCOPE includes a small set of open-ended prompts intended to elicit narrative expression and creative reasoning (Facet 8: Creativity & Innovation). These prompts are qualitatively important because they test whether persona conditioning preserves *style and narrative structure* rather than only discrete response choices. Following prior parameterized creativity evaluation work (Venkit et al., 2025b; Chakrabarty et al., 2024), we quantify differences between human-authored and persona-conditioned LLM narratives along four computational axes inspired by Torrance-style creativity dimensions (flexibility, originality, elaboration, and coherence).

Persona Paradigm	Description & Illustrative Example	Representative Work
Text-Based Descriptions	*Short free-text persona summaries used to condition model behavior.* **Example:** "I am a friendly and curious person who enjoys learning new things and helping others. I value honesty and creativity, and I like discussing technology and everyday life."	Ge et al. (2024)
Sociodemographic Inputs	*Personas defined as structured demographic attribute vectors, often aligned with census statistics.* **Example:** Age: 34; Gender: Female; Race: Hispanic; Education: Bachelor's degree; Income: $50k–$75k; Region: Southwest U.S.	Meyer & Corneil (2025)
Demographic + Identity Signals	*Extends demographic personas with explicit self-identity or cultural identification statements.* **Example:** Age: 29; Gender: Male; Race: Asian American. Identity: "I see myself as both American and strongly connected to my family's cultural traditions. I value collaboration and community belonging."	Lee et al. (2025)
Interview-Based Personas	*Personas derived from interview transcripts or long-form questionnaires capturing lived experience.* **Example:** "I work as a mid-level software engineer maintaining legacy systems. I prefer stability but feel drawn to more creative work. I avoid conflict but speak up on technical issues."	Park et al. (2023)
SCOPE (Ours)	*Personas constructed from structured sociopsychological facets with explicit conditioning and evaluation separation.* **Example:** Demographics (32-year-old Black woman, urban Midwest); Sociodemographic behavior (high civic engagement); Personality traits (high conscientiousness, moderate openness); Identity narrative ("I view my career as a way to support my community"). Evaluation facets (values, behavioral patterns, professional identity, creativity) held out.	*This work*

Table 12: Illustrative persona examples across major persona construction paradigms. Examples are schematic and intended to highlight differences in representational structure.

Demographic Factor	Correlation (r)
Race	+0.110
Political Orientation	+0.090
Country of Residence	+0.082
Occupation	+0.072
Relationship Status	+0.046
Income Satisfaction	+0.044
Religion	+0.028
Nationality	+0.027
Income Category	+0.024
Gender	+0.023
Disability / Health Conditions	+0.005
Age	+0.003
Education Level	+0.002

Table 13: Correlation between individual demographic factors and human response similarity of our human baseline dataset. Values reflect Pearson correlation between demographic similarity and response similarity when considering each factor independently.

B.1 Creativity metric definitions

Let a group of responses (e.g., a persona case, model, or question group) contain n narratives, and let e_i denote the Sentence-BERT embedding for narrative i.

Semantic Diversity (Flexibility). We measure within-group thematic breadth as the average pairwise semantic distance:

$$\text{Diversity} = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left(1 - \cos(e_i, e_j)\right). \tag{8}$$

Higher values indicate broader variation in themes and expression.

Semantic Novelty (Originality). We measure how much a group's average internal distance deviates from the corpus norm:

$$\text{Novelty} = 2 \times |d_{\text{group}} - d_{\text{corpus}}|, \tag{9}$$

where d_{group} is the group's average pairwise semantic distance and d_{corpus} is the corpus-wide average distance. Higher values indicate greater deviation from expected narrative patterns.

Semantic Complexity (Elaboration). We operationalize elaboration via a composite of lexical rarity (TF-IDF) and semantic spread (Word2Vec):

$$\text{C}(s) = 0.5 \times \frac{\text{C}_{\text{TFIDF}}(s)}{\max(\text{C}_{\text{TFIDF}})} + 0.5 \times \frac{\text{C}_{\text{W2V}}(s)}{\max(\text{C}_{\text{W2V}})}. \tag{10}$$

Higher values indicate more intricate narratives in both vocabulary and concept dispersion.

Surprisal (Narrative Coherence). We measure narrative coherence by the average semantic distance between consecutive sentence embeddings within each response:

$$\text{Surprisal} = \frac{2}{n-1} \sum_{i=2}^{n} \left(1 - \cos(e_{i-1}, e_i)\right), \tag{11}$$

where here e_i denotes the embedding of the i-th sentence. Lower surprisal corresponds to smoother semantic progression (more coherent flow).

B.2 Experimental setup

We evaluate creativity on six open-ended SCOPE prompts (Facet 8), comparing human-written responses to LLM-generated responses conditioned on the same persona construction cases used in the main study. We report aggregate comparisons (Human vs. AI) and case-level patterns using the four metrics above.

B.3 Human vs. AI: aggregate findings

Table 17 summarizes aggregate creativity gaps. Consistent with the broader theme of this paper, *persona structure affects behavioral outputs*, we observe that open-ended narrative behavior differs from structured answering behavior: LLM outputs tend to be more elaborative and more deviant from the corpus norm, but less varied across personas and less coherent in narrative flow.

B.4 Persona-case effects: a creativity–constraint trade-off

To understand how persona construction affects narrative-style realism, we compare persona cases by measuring distance from human creativity patterns across the four normalized metrics (equal weight). Interestingly, the most human-like creativity behavior is achieved by **Case 1 (Demographics Only)** (Table 18), suggesting a "creativity–constraint trade-off": richer persona scaffolds that improve structured behavioral fidelity (Sections 4–4.3) may over-constrain open-ended narrative generation and reduce natural variation. This complements our main finding that persona quality is facet-dependent: the optimal conditioning strategy for *structured behavioral prediction* is not necessarily optimal for *open-ended expressive tasks*.

Interpretation. These results indicate that (i) **demographic-only personas** can preserve a surprisingly human-like balance of novelty and diversity in unconstrained narrative tasks, while (ii) **richer sociopsychological scaffolds** (full conditioning and long summaries) tend to increase elaboration but may compress persona-to-persona variability, reducing diversity. Practically, this suggests a deployment guideline: for *creative, narrative, or story-style user simulation*, minimal persona constraints may produce more human-like stylistic variation; for *structured behavior prediction and bias-sensitive simulation*, multi-facet grounding remains preferable (Sections 4–4.3).

SCOPE Case	Description & Synthetic Persona Example (Illustrative)	Included Facets
Case 0: No Persona	*Baseline: no persona conditioning text is provided to the model.* **Example:** (No persona preamble; model answers questions with default behavior.)	None
Case 1: Demographics Only	*Structured demographic attribute vector.* **Example:** **Select Your Age:** 40–44; **Gender:** Male; **Education:** Associate degree; **Occupation:** Logistics Supervisor; **Country:** USA; **Race/Ethnicity:** Black; **Nationality:** American; **Relationship:** Married; **Political Orientation:** Moderate; **Income Category:** $50k–$75k; **Religion:** Christian (non-denominational).	Demographics
Case 2: Demographics + Identity Narratives	*Demographics plus short interview-style identity narratives (personal + professional) that contextualize lived experience.* **Example:** **(Demographics)** 27–29, Female, Bachelor's, USA, Latina, single, left-leaning. **Professional Identity (snippet):** "I'm a public health coordinator. My day is split between community outreach, planning clinics, and writing reports. I measure success by whether residents actually show up and feel respected." **Personal Identity (snippet):** "I grew up translating for my parents, which made me independent early. A turning point was choosing a local college so I could support my family while studying."	Demographics + Professional Identity + Personal Identity
Case 3: Demographics + Traits	*Demographics plus structured traits items and goal statements (Likert-style), emphasizing motivational priorities.* **Example:** **(Demographics)** 33–39, Nonbinary, Master's, USA, White, partnered, center-left. **Traits/Goals (snippet):** "Helping people around me"=5; "Being rich"=2; "Equal opportunities"=5; "Taking risks/adventures"=3; "Secure surroundings"=4; "Thinking up new ideas/being creative"=4. **Short-term goals:** "Pay down debt; build healthier routines; deepen friendships." **Long-term goals:** "Financial stability; meaningful work; more freedom/flexibility."	Demographics + Traits/Goals
Case 4: Full Conditioning	*Combined structured persona: demographics + traits/goals + professional identity + personal identity narratives.* **Example:** **(Demographics)** 50–54, Male, Bachelor's, USA, Asian American, divorced, moderate. **Traits/Goals (snippet):** "Environmental protection"=5; "Tradition"=2; "Achievement/recognition"=3; "Independence"=5. **Professional Identity (snippet):** "I manage IT operations at a mid-sized hospital. The most time-consuming part is incident response and coordinating vendors. I worry about outages and compliance." **Personal Identity (snippet):** "I moved often as a kid, which made me adaptable. A crossroads moment was leaving a stable job to retrain in tech after my first layoff."	Demographics + Traits/Goals + Professional Identity + Personal Identity
Case 5c: LLM Narrative Summary (1000-1500 words)	*Single first-person narrative persona written by the model (compressed, fluent biography-style).* **Example (shortened):** "I'm a 38-year-old single father living in the Midwest. I've built my career in customer support and learned to stay calm under pressure. I care about fairness and try to treat people with respect, even when politics get heated. Lately, I've been thinking about stability, saving more, and finding work that feels meaningful rather than just busy."	AI Narrative Only
Case 6: LLM Facet Completion	*Model is given partial inputs and instructed to infer missing facets to complete a persona profile (ablations vary by what is provided).* **Example (given only demographics):** "Age 22–24, Male, some college, USA, Asian." *Model infers:* plausible values, goals, identity narrative, and work context consistent with the partial seed.	Varies (partial seed + inferred facets)
Case 7: Non-Demographic Persona	*Persona excludes demographics entirely; conditioning uses non-demographic signals (e.g., identity-only, traits-only, or traits+identity).* **Example (traits+identity only):** **Identity (snippet):** "I see myself as a caretaker in my community, but I'm private about my personal life. I'm drawn to groups where people look out for each other." **Traits/Goals (snippet):** "Equality"=5; "Loyalty to friends"=5; "Tradition"=2; "Independence"=4; "Creativity"=3. **Long-term goals:** "Build a stable life; contribute to others; keep autonomy."	Traits and/or Identity (no demographics)

Table 14: SCOPE persona **case** examples used in our experiments. All examples are **synthetic** and included to illustrate the **format and information structure** of each case.

		GPT 4o				Claude-3.5-Sonnet				Gemini 2.0 Flash				Gemini-2.5-Pro-Thinking			
Case	**Description**	r	Acc. (%)	Bias Δ	Bias (%)	r	Acc. (%)	Bias Δ	Bias (%)	r	Acc. (%)	Bias Δ	Bias (%)	r	Acc. (%)	Bias Δ	Bias (%)
Case_1	Demographics Only	0.624	35.07	0.134	101.23	0.627	36.00	0.153	115.666	0.596	35.32	0.071	54.001	0.588	29.78	0.074	55.646
Case_2	Demographics + Narratives	0.642	37.74	0.011	8.298	0.656	38.56	0.012	8.791	0.609	33.58	-0.101	-76.79	0.616	27.58	-0.016	-12.245
Case_3	Demographics + Traits	0.656	38.60	0.100	75.499	0.650	39.60	0.055	41.287	0.600	36.71	0.012	9.087	0.614	22.56	-0.042	-31.64
Case_4	Full Information	0.667	39.67	-0.008	-6.401	0.663	39.98	0.024	17.909	0.614	35.51	-0.064	-48.671	0.633	25.63	-0.016	-11.74
Case_5a	AI Summary (300–500w)	0.645	37.90	-0.040	-30.373	0.642	37.95	-0.036	-27.367	0.602	34.44	-0.077	-58.009	0.608	30.34	-0.103	-78.201
Case_5b	AI Summary (500–1000w)	0.645	38.12	-0.039	-29.80	0.639	37.74	-0.058	-43.601	0.591	32.98	-0.103	-78.321	0.618	31.74	-0.114	-86.013
Case_5c	AI Summary (1000–1500w)	0.649	38.07	-0.051	-38.42	0.648	38.33	-0.055	-41.981	0.591	31.81	-0.117	-88.529	0.620	32.13	-0.120	-90.75
Case_5d	AI Summary (No Limit)	0.646	38.26	-0.041	-31.116	0.638	37.92	-0.056	-42.387	0.594	32.17	-0.096	-72.49	0.614	32.13	-0.111	-84.019
Case_6a	AI (Given Demography only)	0.584	25.53	0.079	59.915	0.460	27.08	0.059	44.782	0.642	19.78	-0.008	-5.976	0.470	24.66	0.011	8.499
Case_6b	AI (Given Demography + Narratives)	0.619	25.55	-0.034	-25.81	0.435	30.18	0.029	21.698	0.649	20.01	-0.041	-30.71	0.517	25.43	-0.065	-49.244
Case_6c	AI (Given Demography + Traits)	0.606	27.86	0.016	12.249	0.462	30.15	0.009	6.797	0.681	21.32	-0.025	-19.097	0.496	27.30	-0.066	-50.171
Case_6d	AI (Given Identity alone)	0.572	39.72	-0.062	-54.69	0.588	41.36	-0.054	-48.15	0.544	39.29	-0.087	-77.65	0.578	40.61	-0.094	-83.38
Case_6e	AI (Given Traits alone)	0.588	41.65	-0.080	-71.43	0.592	41.45	-0.071	-63.02	0.544	40.49	-0.101	-90.04	0.577	40.37	-0.099	-87.90
Case_6f	AI (Given Identity + Traits)	0.608	43.42	-0.046	-40.40	0.614	43.15	-0.038	-33.91	0.561	40.76	-0.079	-70.22	0.586	42.43	-0.082	-72.79
Case_7a	Identity Only	0.633	37.44	-0.086	-65.16	0.649	38.26	-0.072	-54.323	0.601	35.75	-0.101	-76.547	0.604	30.22	-0.123	-93.316
Case_7b	Traits Only	0.624	37.95	-0.094	-71.092	0.639	38.92	-0.088	-66.353	0.595	36.28	-0.121	-91.436	0.595	34.27	-0.115	-87.329
Case_7c	Traits+Identity	0.658	39.68	-0.074	-56.354	0.662	40.29	-0.053	-40.485	0.612	37.86	-0.099	-74.563	0.636	33.89	-0.101	-76.371
AVERAGE	—	0.636	35.53			0.605	36.50			0.613	31.68			0.588	29.12		

Table 15: Complete synthetic persona results (Cases 1–7) for GPT-4o, Claude-3.5-Sonnet, Gemini 2.0 Flash, and Gemini-2.5-Pro-Thinking. We report Pearson correlation with human responses (r), answer accuracy (Acc.), and bias accentuation (Bias Δ and Bias (%) relative to the human baseline).

		DeepSeek R1				GPT 5.1				Qwen3 480b			
Case	**Description**	r	Acc. (%)	Bias Δ	Bias (%)	r	Acc. (%)	Bias Δ	Bias (%)	r	Acc. (%)	Bias Δ	Bias (%)
Case_1	Demographics Only	0.583	31.40	0.005	0.046	0.593	29.30	0.116	1.082	0.545	21.40	0.105	0.977
Case_2	Demographics + Narrative	0.609	35.40	-0.022	-0.207	0.615	31.80	0.035	0.322	0.572	20.90	-0.020	-0.190
Case_3	Demographics + Traits	0.602	35.10	-0.011	-0.107	0.591	29.10	0.050	0.469	0.553	19.10	-0.033	-0.304
Case_4	Full Information	0.625	37.20	-0.039	-0.363	0.621	32.40	0.000	0.003	0.586	21.70	-0.020	-0.188
Case_5a	AI Summary (300–500w)	0.584	34.90	-0.062	-0.575	0.596	33.30	-0.017	-0.160	0.546	22.20	-0.046	-0.427
Case_5b	AI Summary (500–1000w)	0.594	36.00	-0.060	-0.563	0.599	32.70	-0.015	-0.144	0.540	21.90	-0.029	-0.274
Case_5c	AI Summary (1000–1500w)	0.591	36.00	-0.053	-0.490	0.602	32.50	-0.048	-0.451	0.556	22.60	-0.024	-0.221
Case_5d	AI Summary (No Limit)	0.589	34.70	-0.071	-0.658	0.597	31.70	-0.048	-0.442	0.552	22.00	-0.057	-0.533
Case_6a	AI (Given Demography only)	0.539	23.61	0.039	29.486	0.543	21.28	0.032	24.124	0.502	14.47	0.037	28.145
Case_6b	AI (Given Demography + Narratives)	0.572	23.63	-0.030	-23.118	0.576	21.29	-0.025	-18.915	0.532	14.48	-0.029	-22.067
Case_6c	AI (Given Demography + Traits)	0.560	25.76	-0.018	-13.811	0.564	23.22	-0.015	-11.300	0.521	15.79	-0.017	-13.183
Case_6d	AI (Given Identity alone)	0.528	36.73	-0.082	-72.564	0.532	33.10	-0.067	-59.371	0.492	22.51	-0.078	-69.266
Case_6e	AI (Given Traits alone)	0.543	38.52	-0.097	-85.907	0.547	34.71	-0.079	-70.288	0.506	23.61	-0.092	-82.002
Case_6f	AI (Given Identity + Traits)	0.561	40.15	-0.067	-59.763	0.566	36.19	-0.055	-48.897	0.523	24.61	-0.064	-57.046
Case_7a	Narrative Only	0.584	34.62	-0.105	-79.570	0.589	31.20	-0.086	-65.103	0.544	21.22	-0.100	-75.953
Case_7b	Traits Only	0.576	35.10	-0.115	-86.958	0.581	31.63	-0.094	-71.147	0.537	21.51	-0.110	-83.005
Case_7c	Traits+Identity	0.608	36.70	-0.090	-68.138	0.612	33.07	-0.074	-55.749	0.566	22.49	-0.086	-65.040
AVERAGE	—	0.579	33.85			0.584	30.50			0.540	20.73		

Table 16: Complete synthetic persona results (Cases 1–7) for DeepSeek R1, GPT 5.1, and Qwen3 480b. Metrics are defined as in Table 15.

Metric	Human	AI Persona	Direction
Semantic Complexity ↑	0.6933	0.7924	AI higher (more elaboration)
Surprisal (Coherence) ↓	0.9571	1.1384	AI worse (more semantic jumps)
Semantic Diversity ↑	0.1910	0.1685	Human higher (more variation)
Semantic Novelty ↑	0.0181	0.0329	AI higher (more deviation)

Table 17: Aggregate creativity and narrative-style comparison between human-authored responses and persona-conditioned across all 7 model responses across the creativity prompts. Arrows indicate preferred direction for *human-like* narrative quality: higher diversity/complexity/novelty, and lower surprisal.

Rank	Case	Distance to Human (lower is better)
1	Case_1 (Demographics only)	0.5473
2	Case_2D (Narrative only; no demographics)	0.5763
3	Case_2 (Demographics + identity)	0.5868
4	Case_4 (Full conditioning)	0.6010
5	Case_2S (Narrative augmented)	0.6181

Table 18: Top creativity-aligned persona cases under a normalized, equal-weight distance over semantic complexity, surprisal, semantic diversity, and semantic novelty across all 7 models. Lower distance indicates closer alignment to human creativity patterns.