The Need for a Socially-Grounded Persona Framework for User Simulation

Anonymous ACL submission

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 values 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 and 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., 2023). As summarized in Table 1, 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 and Corneil, 2025), and related systems treat demographics or brief identity cues as proxies for human individuality (Kane and Schubert, 2023). Table 2 further shows that these frameworks capture only limited slices of sociopsychological structure, typically omitting psychological traits, values, and behavioral patterns that social science research identify as central drivers of behavior (Turner and Reynolds, 2012; Costa Jr and McCrae, 2000). This reliance on demographic-centric abstractions persists despite evidence that demography 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 **o**f **P**ersona **E**valuation), a human-grounded framework for constructing and evaluating synthetic personas. Motivated by the limitations of existing paradigms (Table 1), 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 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. First, demographic-only personas constitute a structural bottleneck: although demographic similarity explains only $\sim$1.5% of behavioral variance among humans, demographic-only prompting often more than doubles this signal in model behavior, indicating systematic overgeneralization. Second, adding sociopsychological grounding steadily improves behavioral alignment while reducing demographic over-accentuation, with full SCOPE conditioning achieving the strongest overall correlation. Third, non-demographic personas based on values and identity alone can match or exceed fully conditioned personas while substantially reducing demographic bias. Finally, longer or more fluent model-written summaries do not substitute for structured, human-grounded facets, demonstrating that persona *structure* matters as much as information quantity. We further validate these findings on SimBench (Hu et al., 2025), a large external benchmark of social and behavioral questions, where SCOPE personas outperform default prompting and NVIDIA Nemotron personas, and SCOPE-based augmentation improves Nemotron's performance. By grounding persona construction and evaluation in human data, SCOPE provides a foundation for building personas that are more behaviorally realistic, less demographically stereotyped, and better suited for sociotechnical NLP systems (Venkit et al., 2025a). SCOPE also acts as an augmentation pipeline to existing frameworks making it immediately actionable for practitioners who rely on large-scale persona generation pipelines.

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 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 et al., 2001; Turner and Reynolds, 2012). Personality psychology further demonstrates that stable traits, most prominently the Big Five to predict consistent patterns of cognition, affect, and behavior across contexts (Costa Jr and McCrae, 2000; Costa and 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 when structured patterns across situations rather than single responses in isolation (Mischel and 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 and 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.

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

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 and 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 1: Overview of major persona construction frameworks in current LLM research. Example of each of the framework is present in Table 10–11 in the Appendix.

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., 2025). 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 for conversational agents, recommender systems, and social simulation, 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**. SCOPE is a sociopsychologically informed framework for constructing rich, multidimensional persona representations.

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 8). 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 1 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.4).

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 8. This enables direct comparison with real human responses, unlike prior work

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

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.3-A.4[3]).

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) (Smith et al., 2023; 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., 2024; Christoforou et al., 2024). GPTZero was used as a coarse AI-authorship filter, supplemented by manual review (Brown and Jensen, 2023; gpt, 2025). Participants were compensated $50; 124 completed the survey with verified human-authored responses, yielding 17,484 total responses (124 × 141). 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), 2.4% 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, question-type distributions, response-option frequencies are in Appendix A.4.

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 8), 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. Baseline case that contains no persona informaiton. This case shows the improvement of using personas.

Case 1: Demography-Only. Personas are constructed from basic sociodemographic attributes. This resembles how most works construct personas.

Case 2: Demography + Narratives. Adds personal and professional identity narratives on top of demographics.

Case 3: Demography + Traits. Demographics with Big Five traits derived from the BFI-2-S.

**Case 4: Full Conditioning (Demography +

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

[3] The complete survey is in the Supplementary Materials

Traits + Narratives + Sociodemographic Behavior). Uses all conditioning facets..

We additionally introduce three **LLM-centric** variants to mirror current practices in automated persona generation:

Case 5: LLM Narrative Summaries. The LLM receives a condensed persona summary generated *by the model itself* from the full SCOPE profile. This compresses response into a single narrative persona as seen in Nemotron.

Case 6: LLM Facet Completion. The model is given only partial information (e.g., demographics or demographics+identity) and instructed to infer the missing facets. Multiple ablation of what information is used will create case 6 subsets.

Case 7: Non-Demographic Personas. Personas that omit demographics entirely (e.g., identity-only or values+identity personas). These variants evaluate whether non-demographic cues alone can 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 11 in the Appendix.

4 Evaluation Framework

Our evaluation intends to address 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*. This distinction is essential as humans rarely provide identical responses, but they often display stable patterns across related questions (Kambhatla et al., 2022; Venkit et al., 2025b). Given that humans themselves disagree item-by-item, 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 therefore 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 first analyze these metrics on our SCOPE persona ablations (Cases 1–7), then evaluate generalization on SimBench (Hu et al., 2025), 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 golden standard text set consisting of questions from *sociodemographic behaviours*, *personal values and motivations*, and *behavioral patterns* to understand how each case is able to predict and replicate these facets. This leads to a total MCQ set of 72 questions which is used as the first baseline of evaluation.

Correlation. We measure behavioral alignment using the Pearson correlation coefficient

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

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:

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

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 1 shows the correlation results across all the models used and Table 3 and 4 summarizes GPT-4o performance[4] across the most informative persona cases; the full 7-model breakdown appears in Table 12 – 13. Three trends emerge across all the model performance:

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

[4]We show GPT4o results as they are the better performing of all the models. The complete result is in the Appendix.

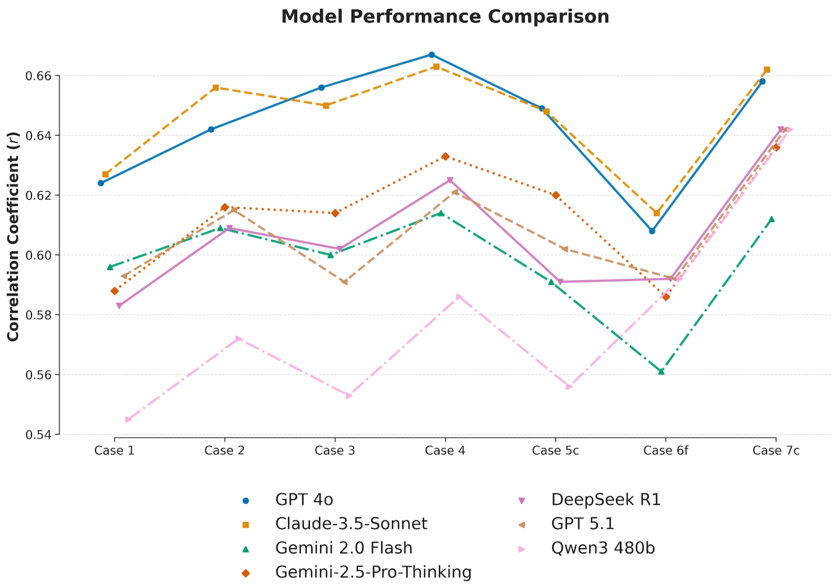

Figure 1: Line graph denotes the correlation score of all seven LLM modes.

identity narratives (Case 2), values (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 values and identity (Case 7c) match or exceed the best full-information case in correlation while avoiding direct use of demographic attributes.

IV. Personas grounded in non-sociodemographic information perform better. The greater finding show that to understand social behaviours, sociodemographic information is not the most important element of a persona. We see that for Case 6 iterations, where personas are created using either or both values or identity information tend to perform better, *in regards to accuracy*, than the ones generated by sociodemographic information.

V. Socially grounded personas are able to simulate sociodemographic and behavioral patters. Table 4 also shows how amongst all tasks, personas are able to replicate well sociodemographic behaviours and personality traits. The results also show that unique day to day behaviours are capturable but harder to replicate. The results show that personas can be used well for market specific understanding of a group pattern as compared to individualistic preferences.

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: + Identity narratives	0.642	37.74	8.30
3: + Values & motivations	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+values	0.608	43.42	-40.40
7c: Values + identity only	0.658	39.68	-56.35

Table 3: 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.

Across models, we see similar trends: richer, multi-facet personas consistently yield higher correlations than demographic-only cases, and non-demographic SCOPE personas (Case 7c) match or exceed full conditioning while remaining more conservative with respect to demographic bias. We additionaly also explore open ended questions in Appendix B.4 where we use framework defined by Venkit et al. (2025b); Chakrabarty et al. (2024) to evaluate writing style and diversity of text based answers. Our findings show how writing styles are very facet dependent and as a task is sensitive.

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, Lati-

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 4: A snippet of facet-level accuracy scores across persona construction cases for Socio.: *sociodemographic behavior*, Behavioural: *behavioral patterns*, and Traits: *personality traits*.

Persona Source	Case Description	Accuracy
SCOPE (Ours)	4. Full conditioning	**0.584**
	2. Demo. + Identity	0.557
	6a. AI profile	0.553
	3. Demo. + values	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 5: Task-based evaluation on **SimBench**.We report accuracy using GPT-4o personas under different construction strategies.

noBarometro, OpinionQA). This provides a task-based understanding of how our persona framework performs well in behavioral and social tasks. We align SimBench questionnaire items with our personas created to retain 441 questions whose wording, response scale, and sociotechnical context are compatible with the sociodemographic groups present in our collected data and personas.

We created GPT-4o personas[5] for each human participant and persona case, then prompted the model to answer the **441 SimBench questions** in character. For each SimBench question k we aggregate the human responses for the matching demographic group into an empirical distribution p_k, and treat the persona answer as a degenerate distribution q_k. We then compute **majority-vote accuracy**, i.e., whether the persona selects the most frequent human response for that group (within 3 turns).

We also compare against personas sampled from the NVIDIA Nemotron USA persona dataset (Meyer and Corneil, 2025), using their full textual persona descriptions and the same GPT-4o backend. Illustrated examples of Simbench and Nemotron are in Appendix A.7. SCOPE personas that include sociopsychological facets (Cases 2–4 and 7c) outperform Nemotron personas across the 441 SimBench items, while demographic-only SCOPE personas perform comparably to or slightly better than Nemotron. This supports our central claim: *one-dimensional sociodemographic personas are not sufficient even on external benchmarks, whereas multi-facet SCOPE personas does better to real-world tasks*. This finding is greatly also seen where Nemotron personas, augmented with SCOPE facets, becomes the best performing variant within the Nemotron persona cases. This shows that *SCOPE framework can be used as an augmentation of existing personas to increase their social and behavioral information and attributes to obtain strongly similar predictive behaviors.*

[5] as GPT4o is the current high-performing model

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 beyond empirical human baselines.

Human baseline: We estimate the predictive strength of demographics on behavioral similarity using $N_{\text{human}} = 117$ participants, 13 demographic variables, and 134 evaluation questions. For each participant pair (i, j), we compute demographic similarity d_{ij} and response similarity s_{ij}^{human} (Pearson correlation), and define the resulting demographic influence coefficient as:

$$r^{\text{human}} = \operatorname{corr}\big(d_{ij},\, s_{ij}^{\text{human}}\big). \tag{4}$$

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 provides a conservative baseline 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:

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

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

$$\Delta r_{m,c} = r^{\text{AI}}_{m,c} - r^{\text{human}}, \quad (6)$$

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

Positive values show persona makes demographically similar individuals *too similar* (over-accentuation), while negative values indicate that the persona *underplays* demographic structure.

Results across cases and models: The results in Table 3 and Appendix Tables 12–13 show a consistent pattern. First, **demographic-only prompting strongly over-accentuates demographic similarity**. For GPT-4o, Case 1 yields $\text{Bias}\% = 101.23$ (Table 3), i.e., more than doubling the demographic signal observed in the human baseline. Claude-3.5-Sonnet exhibits an even larger effect in Case 1 ($\text{Bias}\% = 115.67$; Table 12). This indicates that, when instructed with only demographic variables, models tend to collapse behavior toward demographic stereotypes.

Second, **adding sociopsychological grounding substantially reduces demographic accentuation**, though the magnitude varies by facet. For GPT-4o, Case 2 (demographics + identity) reduces bias to near-baseline ($\text{Bias}\% = 8.30$), whereas Case 3 (demographics + values) still shows non-trivial over-accentuation ($\text{Bias}\% = 75.50$), suggesting that values alone do not necessarily prevent demographic clustering when demographics remain explicit inputs. In contrast, Case 4 (full conditioning) slightly *under-shoots* the human baseline for GPT-4o ($\text{Bias}\% = -6.40$), indicating that SCOPE can preserve behavioral alignment.

Third, **non-demographic personas are the most conservative with respect to demographic bias**. For GPT-4o, Case 7c (values + identity, no demographics) yields $\text{Bias}\% = -56.35$ while maintaining high correlation with human response structure ($\bar{r} = 0.658$; Table 3). A similar pattern appears across other models in Appendix Tables 12–13, where identity- and values-driven variants tend to be negative (or substantially reduced) in Bias% relative to demographic-only variants.

5 Discussion and Conclusion

Our goal is to re-examine how synthetic personas are constructed, evaluated, and interpreted in NLP/AI. Although personas are widely used for evaluation, and social simulation, our results reveal a systematic mismatch between demographic-centric persona design and the structure of real human behavior, with implications for persona modeling, evaluation, and bias-aware design.

Demographic-centric personas are a structural bottleneck. Personas built solely from sociodemographic attributes (Case 1) consistently underperform across models, showing lower behavioral correlation, lower accuracy, and the strongest demographic bias amplification (Tables 3, 12, 13), a pattern that also holds on SimBench. In human data, demographic similarity explains only $\sim$1.5

Sociopsychological grounding improves alignment. Adding values, traits, identity narratives, and behavioral signals steadily improves alignment across models, with full conditioning (Case 4) achieving the highest correlation while reducing demographic bias. These gains are not driven by verbosity: AI-generated summaries (Case 5) are often longer yet underperform structured, human-grounded representations, indicating that persona *structure* matters as much as information quantity. Consistently, non-demographic personas based on values and identity (Case 7c) match or exceed full conditioning in correlation while substantially reducing bias, showing that demographics are neither necessary nor sufficient.

Bias emerges from persona design choices. Demographic-only personas systematically over-accentuate demographic clustering, effectively encoding stereotypes into simulated behavior. In contrast, SCOPE reduces bias without sacrificing realism: full conditioning closely matches the human baseline, while values+identity personas are the most conservative with respect to demographic amplification while maintaining strong alignment. This supports a practical principle: bias reduction is best achieved by replacing coarse demographic proxies with psychologically meaningful structure, rather than removing structure altogether.

SCOPE as an augmentation framework. Rather than replacing existing persona systems, SCOPE functions as a modular augmentation layer. Applied to Nemotron personas, SCOPE-based augmentation consistently improves accuracy and alignment on SimBench, outperforming Nemotron's native representations. This positions SCOPE not merely as a dataset or benchmark, but as a general-purpose framework for enriching persona representations without abandoning existing persona corpora.

Ethics Statement

We collected a two-hour sociopsychological questionnaire from U.S.-based adult participants under review by the institution's Ethics Office. Participants provided informed consent and were compensated $50 for completion. To reduce privacy risk, we avoided collecting and removed direct identifiers (e.g., legal names, email addresses, phone numbers, street addresses). Free-text responses were stored under anonymized participant IDs and screened to remove accidental disclosure of sensitive personal information. Because persona conditioning can amplify stereotypes, we explicitly measure demographic accentuation (Section 4.3) and report persona variants that achieve strong behavioral alignment while reducing demographic clustering. We release only the Nemotron augmented data and AI profile materials, and report aggregate statistics to omit any re-identification risk. *This can be found in our Supplementary Materials.*

The full study protocol underwent internal ethical review prior to deployment. The review examined risks associated with collecting sociopsychological and narrative data, potential re-identification vectors, and the fairness implications of using such data for AI persona modeling. To protect participants, we intentionally avoided collecting direct personal identifiers such as names, email addresses, street addresses, or phone numbers. All responses were stored under anonymized participant identifiers, and raw text responses were screened for accidental disclosure of sensitive personal information.

No sensitive protected health information or financial identifiers were collected. The anonymized dataset used for modeling contains only aggregated demographic categories and de-identified narrative text. Access to the full dataset is restricted to the research team, and all analyses reported in this paper were conducted on the de-identified corpus. These procedures were designed to provide strong privacy protection while enabling rigorous sociotechnical research on persona-grounded AI systems. In regards to AI usage in the paper, we used AI-assisted tools solely for grammatical and stylistic corrections of our writing. All scientific content, analysis, and conclusions are the authors' own.

Limitations

Our study acknowledges the following limitations. First, the participant pool is U.S.-based; while we target Census-like proportions and oversample several underrepresented groups, behavioral norms and identity narratives are culturally contingent and may not transfer globally. Second, our evaluation uses a fixed survey instrument; although it spans eight primary facets and multiple response modalities, it cannot cover all drivers of behavior (e.g., longitudinal life events, situational stressors, or fine-grained local context). Third, correlation-based structural evaluation captures relative response patterns but does not guarantee causal fidelity or calibration; two systems can correlate similarly while differing in absolute distributions or rationale quality. Finally, our data-quality controls for AI-authored responses are imperfect; automated detectors are at best heuristic, and filtering decisions may introduce selection artifacts. Our future work intends to incorporate stronger verification (e.g., synchronous interviews, response-time features, or multi-signal provenance checks) and broader cross-cultural replications.

References

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

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.

Jan Batzner, Volker Stocker, Bingjun Tang, Anusha Natarajan, Qinhao Chen, Stefan Schmid, and Gjergji Kasneci. 2025. 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, pages 343–354.

Manuela Benary, Xing David Wang, Max Schmidt, Dominik Soll, Georg Hilfenhaus, Mani Nassir, Christian Sigler, Maren Knödler, Ulrich Keller, Dieter Beule, and 1 others. 2023. Leveraging large language models for decision support in personalized oncology. *JAMA Network Open*, 6(11):e2343689–e2343689.

David W Brown and Dean Jensen. 2023. Gptzero vs. text tampering: The battle that gptzero wins. *International Society for Technology, Education, and Science*.

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2024. 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*, pages 1–34.

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023. 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)*, pages 1504–1532.

Evgenia Christoforou, Gianluca Demartini, and Jahna Otterbacher. 2024. 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, pages 2097–2103.

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

Paul T Costa Jr and Robert R McCrae. 2000. *Neo Personality Inventory*. American Psychological Association.

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

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

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

Hangzhi Guo, Pranav Narayanan Venkit, Eunchae Jang, Mukund Srinath, Wenbo Zhang, Bonam Mingole, Vipul Gupta, Kush R Varshney, S Shyam Sundar, and Amulya Yadav. 2025. 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, pages 1169–1180.

Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. 2023. Bias runs deep: Implicit reasoning biases in persona-assigned llms. *arXiv preprint arXiv:2311.04892*.

Tiancheng Hu, Joachim Baumann, Lorenzo Lupo, Nigel Collier, Dirk Hovy, and Paul Röttger. 2025. Simbench: Benchmarking the ability of large language models to simulate human behaviors. *arXiv preprint arXiv:2510.17516*.

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

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

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

Benjamin Kane and Lenhart Schubert. 2023. 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*, pages 10998–11016.

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

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

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

Keyeun Lee, Seo Hyeong Kim, Seolhee Lee, Jinsu Eun, Yena Ko, Hayeon Jeon, Esther Hehsun Kim, Seonghye Cho, Soeun Yang, Eun-mee Kim, and 1 others. 2025. 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)*, pages 6971–6991.

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

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

Yev Meyer and Dane Corneil. 2025. Nemotron-Personas-USA: Synthetic personas aligned to real-world distributions.

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

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

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. 2024. Internet & technology. https://www.pewresearch.org/internet/. Accessed 2024.

Haoyi Qiu, Yilun Zhou, Pranav Narayanan Venkit, Kung-Hsiang Huang, Jiaxin Zhang, Nanyun Peng, and Chien-Sheng Wu. 2025. Mmpersuade: A dataset and evaluation framework for multimodal persuasion. *arXiv preprint arXiv:2510.22768*.

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

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

Tom W. Smith, Peter Marsden, Michael Hout, and Jibum Kim. 2023. General social surveys, 1972–2022: Cumulative codebook. https://gss.norc.org.

Henri Tajfel, John Turner, William G Austin, and Stephen Worchel. 2001. An integrative theory of intergroup conflict. *Intergroup relations: Essential readings*, pages 94–109.

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

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

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

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 6: 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.

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.

Simone Zhang, Janet Xu, and A Alvero. 2024. Generative ai meets open-ended survey responses: Participant use of ai and homogenization.

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 Facet-level grounding and coverage.

Table 8 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.2 End-to-end Pipeline.

Figure 3 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,

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

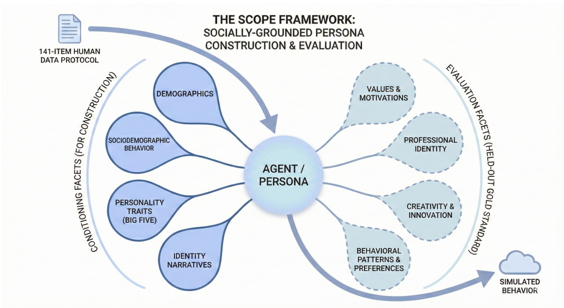

Figure 2: Figure illustrates the facet used for constructing the persona to the ones used for evaluation as a baseline in creating better personas.

and summarized through correlation, accuracy, and bias metrics.

A.3 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. For space, we do not reproduce the full questionnaire; instead, Table 9 provides one illustrative example item per facet along with response formats.

A.4 Survey Statistics

Our survey contains 141 questions across eight facets (Table 8). 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 6–7 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.5 Framework Examples

We show how each of the persona framework differs by showcasing examples of each framework in Table 10. We see how our framework contains more details and information to process behavioural and social tasks better.

A.6 Complete model-level results.

Tables 12 and 13 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 grounding, and the robustness of non-demographic personas, hold consistently across models with very different architectures and training regimes.

A.7 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 and Corneil, 2025) dataset[6] consists of fully synthetic U.S. personas generated to mirror census-grounded demographic distributions and provide diverse, non-

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

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), Smith et al. (2023), 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 and 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 8: 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.

PII persona descriptions.[7] 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*.

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

SimBench (group prompt template + response distribution). SIMBENCH (Hu et al., 2025) 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 primary

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 9: 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.

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:

> **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}`

A.8 From structured behavior to expressive behavior.

While the main paper focuses on structured behavioral alignment (multiple-choice, Likert, and categorical responses), SCOPE also includes a small but important set of open-ended prompts designed to probe narrative expression, creativity, and writing style. These prompts capture aspects of human behavior that are not easily reduced to discrete labels, such as expressive flexibility, originality, and narrative coherence. In the following section, we analyze these open-ended responses using established computational creativity metrics, complementing our correlation-based evaluation and further illustrating how persona structure shapes

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 and 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 10: Illustrative persona examples across major persona construction paradigms. Examples are schematic and intended to highlight differences in representational structure.

model behavior across different response modalities.

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).

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:

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

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:

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

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):

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

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

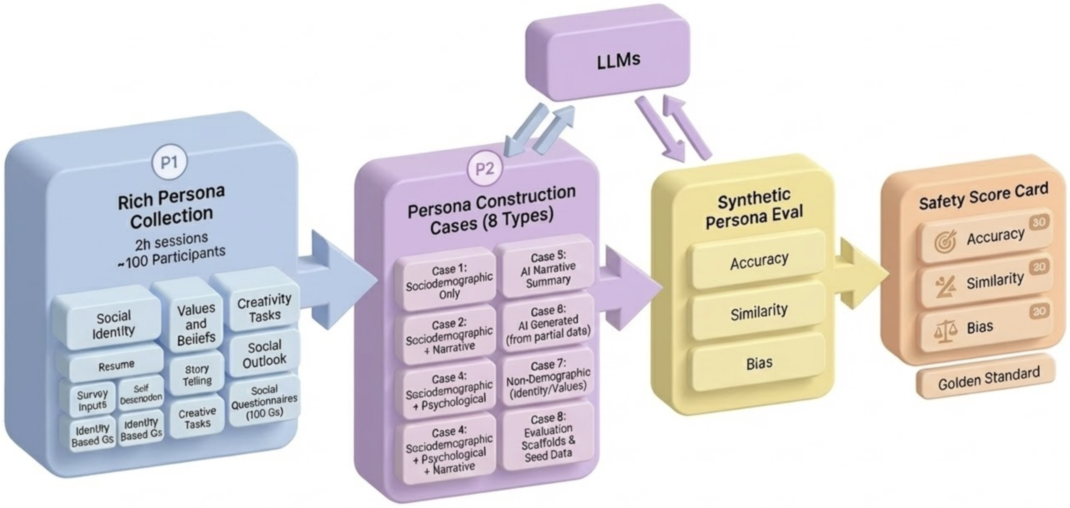

Figure 3: **End-to-end pipeline of our persona construction and evaluation framework.** The process consists of four major stages: (1) **Rich Persona Collection**: a two-hour, 141-item sociopsychological survey capturing eight behavioral facets across 100 participants; (2) **Persona Construction Cases**: eight 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 bias; and (4) **Safety Score Card**: aggregating evaluation metrics into a standardized benchmark for persona quality. This workflow provides a unified, human-grounded framework for constructing and assessing synthetic personas compatible with modern LLM systems.

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

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

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 14 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 15), 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 + Values & Goals	*Demographics plus structured value items and goal statements (Likert-style), emphasizing motivational priorities.* **Example:** **(Demographics)** 33–39, Nonbinary, Master's, USA, White, partnered, center-left. **Values/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 + Values/Goals
Case 4: Full Conditioning	*Combined structured persona: demographics + values/goals + professional identity + personal identity narratives.* **Example:** **(Demographics)** 50–54, Male, Bachelor's, USA, Asian American, divorced, moderate. **Values/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 + Values/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, values-only, or values+identity).* **Example (values+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." **Values/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."	Values and/or Identity (no demographics)

Table 11: 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 + Identity	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 + Values	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 + Identity)	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 + Values)	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 Values 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 + Values)	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	Values 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	Values+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 12: 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 + Identity	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 + Values	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 + Identity)	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 + Values)	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 Values 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 + Values)	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	Identity 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	Values 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	Values+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 13: Complete synthetic persona results (Cases 1–7) for DeepSeek R1, GPT 5.1, and Qwen3 480b. Metrics are defined as in Table 12.

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 14: 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 (Identity only; no demographics)	0.5763
3	Case_2 (Demographics + identity)	0.5868
4	Case_4 (Full conditioning)	0.6010
5	Case_2S (Identity augmented)	0.6181

Table 15: 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.