Parametric Social Identity Injection and Diversification in Public Opinion Simulation

Hexi Wang
DCST, Tsinghua University
Beijing, China
Quancheng Laboratory
Jinan, Shandong, China
whx25@mails.tsinghua.edu.cn

Yujia Zhou*
Quancheng Laboratory
Jinan, Shandong, China
DCST, Tsinghua University
Beijing, China
zhouyujia@mail.tsinghua.edu.cn

Bangde Du
DCST, Tsinghua University
Beijing, China
dbd23@mails.tsinghua.edu.cn

Qingyao Ai*
Quancheng Laboratory
Jinan, Shandong, China
DCST, Tsinghua University
Beijing, China
aiqy@tsinghua.edu.cn

Yiqun Liu
DCST, Tsinghua University
Beijing, China
yiqunliu@tsinghua.edu.cn

Abstract

Large language models (LLMs) have recently been adopted as synthetic agents for public opinion simulation, offering a promising alternative to costly and slow human surveys. Despite their scalability, current LLM-based simulation methods fail to capture social diversity, producing flattened inter-group differences and overly homogeneous responses across demographic groups. We identify this limitation as a Diversity Collapse phenomenon in LLM hidden representations, where distinct social identities become increasingly indistinguishable across layers. Motivated by this observation, we propose Parametric Social Identity Injection (PSII), a general framework that injects explicit, parametric representations of demographic attributes and value orientations directly into intermediate hidden states of LLMs. Unlike prompt-based persona conditioning, PSII enables fine-grained and controllable identity modulation at the representation level. Extensive experiments on the World Values Survey using multiple open-source LLMs show that PSII significantly improves distributional fidelity and diversity, reducing KL divergence to real-world survey data while enhancing overall diversity. This work provides new insights into representation-level control of LLM agents and advances scalable, diversity-aware public opinion simulation.

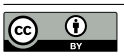
CCS Concepts

• **Computing methodologies** → **Natural language processing**; *Artificial intelligence*; • **Applied computing** → *Sociology*.

Keywords

Agent-based Modeling, Public Opinion Simulation, Social Diversity

*Corresponding authors.



KDD '26, Jeju Island, Republic of Korea

ACM ISBN 979-8-4007-2259-2/2026/08
https://doi.org/10.1145/3770855.3817926

ACM Reference Format:
Hexi Wang, Yujia Zhou, Bangde Du, Qingyao Ai, and Yiqun Liu. 2026. Parametric Social Identity Injection and Diversification in Public Opinion Simulation. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '26), August 09–13, 2026, Jeju Island, Republic of Korea.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3770855.3817926

Resource Availability:
The source code and data used in this paper are publicly available at https://github.com/halsayxi/PSII, with a versioned archival release available at https://doi.org/10.5281/zenodo.20465632.

1 Introduction

Public opinion simulation [13] is critical for quantifying societal attitudes, yet traditional surveys face escalating costs and scalability issues [10, 14, 23, 35]. To address this, recent work explores **agent-based modeling** (ABM) using **large language models** (LLMs) as synthetic respondents. Leveraging LLMs enables efficient, low-cost simulations with several advantages, including but not limited to reduced logistical burdens, flexible experimental scenario design, unlimited follow-ups, and multi-dimensional and controllable experimental populations conditioned on demographic, socioeconomic, or ideological attributes.

While LLM-based opinion simulation approaches demonstrate promising performance on specific tasks, they often share a critical limitation: insufficient diversity in simulated populations. Diversity is crucial for social research and public opinion studies. Even small degrees of population homogenization or representational bias can lead to misleading conclusions about societal dynamics [12, 16, 21, 26, 30, 32]. Yet, as shown in our experiments and previous studies [5, 6, 20, 22, 28], existing LLM-based approaches often produce homogeneous results whose behavioral distributions differ significantly from those of real human subjects. LLM-based approaches to public opinion simulation from previous studies generally can be categorized as direct zero-shot querying [31], persona-based prompting [4, 11, 19, 41, 42], or fine-tuning based alignment [18, 34, 38]. Their limitations in diversity manifest at two

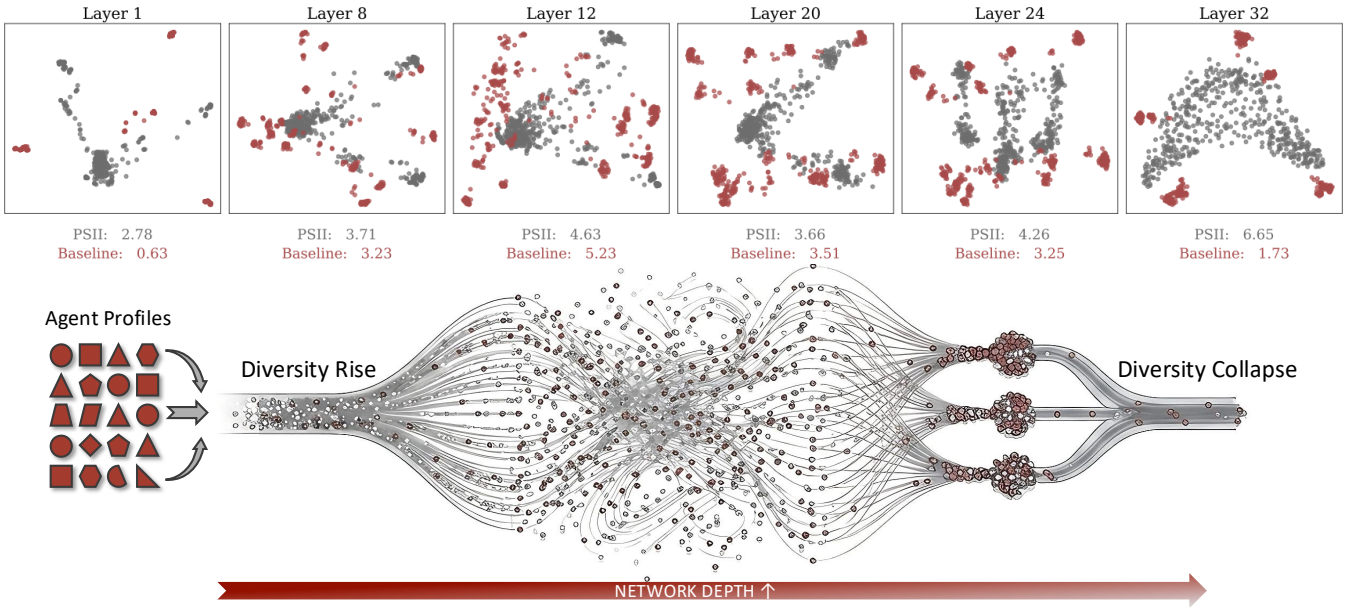


Figure 1: Layer-wise scatter plots of final-token hidden states for 500 simulated agents (top) and an illustration of Diversity Collapse in Transformer hidden states (bottom). In the top panels, red points denote baseline methods and gray points denote PSII-generated agents; the reported scores measure the average spatial dispersion of representations in each layer. The bottom panel depicts the Diversity Collapse phenomenon.

distinct levels. First, the **inter-group diversity**. Standard training objectives for LLMs, such as maximum likelihood estimation with cross-entropy loss, inherently favor the most probable continuations. As a result, minority or low-frequency viewpoints tend to be underrepresented, leading to flattened response distributions in which distinct subpopulations become difficult to distinguish [37]. Second, **intra-group diversity**. When demographic attributes are injected via prompts, identities are treated as fixed explanatory variables that dominate response generation [37]. This method ignores the heterogeneity within groups, exaggerates between-group differences, and reinforces group stereotypes as a consequence.

To understand diversity loss in LLM agents, we analyze **internal hidden-state representations**. We projected final-token hidden states from 500 agents using KPCA to visualize representational diversity across layers (Figure 1). From the red baseline points, our analysis reveals a non-monotonic pattern: lower layers form compact clusters, while intermediate layers spread out, reaching high diversity. Critically, higher layers experience a systematic contraction, collapsing into dense clusters, a phenomenon we term **Diversity Collapse**. The above analysis highlights a fundamental limitation of existing approaches: they lack stable and heterogeneous conditions to guide hidden-state evolution throughout the network. See Appendix E.4 for details.

The analysis highlights that existing approaches lack stable conditions to guide hidden-state evolution. As external inputs, textual prompts are progressively smoothed across layers, making them insufficient for sustaining structured individual differences. Consequently, synthetic agents often possess weakly grounded identity representations. Motivated by these observations, we propose **Parametric Social Identity Injection (PSII)**, which explicitly models social identity within the internal representation space. Similar to Parametric RAG [33], PSII embeds identity information as parametric vectors, including **demographic and value vectors**, directly into hidden states. This enables identity attributes to shape hidden-state trajectories rather than relying on surface prompts. To enhance **inter-group diversity**, PSII introduces stable signals that persist across layers; to preserve **intra-group diversity**, we apply stochastic perturbations to simulate natural variation. Beyond effectiveness, PSII offers three key advantages: **efficiency**, using vectors to modulate identities with minimal storage and no fine-tuning; **reusability**, which forms a modular library of agent identities applicable across various datasets and enables the efficient generation of diverse synthetic populations at a scale proportional to the Cartesian product of attribute dimensions; and **tractability**, enabling precise quantitative analysis via linear algebraic operations.

We evaluate PSII on the World Values Survey (WVS)[1] dataset using multiple open-source LLMs, including Qwen2.5-7B-Instruct and Qwen2.5-14B-Instruct [3], Llama-3.1-8B-Instruct [36], and Mistral-24B-Instruct [25]. Across models and tasks, PSII consistently improves both prediction accuracy with respect to human responses and diversity metrics. Notably, in the same layer-wise visualization (Figure 1, top), the gray points show representations produced by PSII, which exhibit sustained or even increasing diversity in higher layers, effectively counteracting the Diversity Collapse observed in baseline methods.

In summary, this work makes three primary contributions. First, we identify and characterize the Diversity Collapse phenomenon

[1]https://www.worldvaluessurvey.org/wvs.jsp

in the hidden states of LLM-based social simulation agents, which explains why conventional prompt-based methods fail to capture population heterogeneity. Second, we propose Parametric Social Identity Injection (PSII), a principled framework for stable and heterogeneous identity modeling by injecting identity vectors into hidden representations. Third, through systematic experiments on WVS, we demonstrate that PSII significantly improves diversity and distributional fidelity, producing synthetic populations that better reflect real-world heterogeneity.

2 Related Work

2.1 LLM-based Personality Simulation Agents

Early research focused on LLMs' zero-shot capabilities. While direct prompting of models on specific topics is a fundamental baseline, studies show default outputs often exhibit political biases (e.g., left-leaning) and underrepresent marginalized groups [2, 31].

To improve realism, Persona-based Prompting injects demographic attributes (age, gender, race) into prompts [41, 42, 44]. Hwang et al. [19] proposed a framework for aligning with user opinions, demonstrating that persona-based personalized prompting significantly improves prediction accuracy. However, Beck et al. [4] pointed out in their study that Sociodemographic Prompting involves a trade-off between sensitivity and robustness, and simple attribute injection may lead to model stereotyping.

Recent trends shift toward Fine-tuning and Alignment [34, 38]. Compared to general-purpose models, models fine-tuned on specific social survey data exhibit stronger distribution fitting capabilities. For instance, SimVBG simulates complex values via individual backstories [11], while Distribution Shift Alignment (DSA) helps models adapt to context changes [18].

To better mirror real-world demographics, Chen et al. [8] proposed HAG (Hierarchical Demographic Tree-based Agent Generation), utilizing a hierarchical tree structure to generate topic-adaptive agents. Hu et al. [17] introduced Population-Aligned Persona Generation, utilizing importance sampling to reduce population-level biases. Chen et al. [7] further explored precisely regulating model personality traits by controlling specific directions in the activation space, providing a new technical pathway for high-fidelity social simulation.

2.2 Enhancing Diversity and Representativeness

Lack of diversity remains a challenge, where models favor generic opinions over minority voices.

Traditional sampling strategies, such as high-temperature or top-k sampling, can increase randomness but often come at the cost of response coherence [9, 29]. To enhance diversity while maintaining quality, Wong et al. [40] proposed SimpleStrat, leveraging the concept of stratification to guide the model in exploring different solution spaces. Zhang et al. [43] introduced Negatively-Correlated (NC) Sampling, which forces the model to unearth differentiated perspectives by suppressing the probability of already generated opinions, and released the Community Alignment Dataset to support research on pluralistic preferences.

In Prompt Engineering, mechanisms like Step-by-step Recall and Collective-Critique (CCSV) enhance viewpoint coverage and cultural diversity [15, 24]. Multilingual Prompting also serves as an implicit cue to activate embedded cultural knowledge [39]. Finally, research by Abels et al. [1] suggests that relying solely on LLM populations may exacerbate biases. They proposed the concept of Hybrid Human-LLM Crowds, demonstrating that combining human diversity with LLM reasoning capabilities is an effective approach to mitigate bias and enhance collective intelligence.

3 Parametric Social Identity Injection

We propose **Parametric Social Identity Injection (PSII)**, a framework for enhancing both fidelity and diversity in LLM-based public opinion simulation. As illustrated in Figure 2, PSII injects structured identity information, including demographic and value-related features, directly into the model's hidden states, introducing stable and heterogeneous individual differences that guide the generation of synthetic responses.

3.1 Identity Construction

PSII integrates two complementary components to model individual agents: **agent profiles** (prompt level) and **identity vectors** (hidden state level). This dual-level approach addresses the lack of stable personality modeling in conventional LLM simulations.

3.1.1 Agent Profile Description. For each synthetic agent, we construct a semantic agent profile using demographic variables. These variables are converted into descriptive text prompts, providing the model with semantic priors about the agent's demographic context. Formally, let P_i denote the agent profile of agent i, composed of a set of descriptive phrases:

$$P_i = \{d_1, d_2, \ldots, d_M\},$$

where d_j is the textual description corresponding to the j-th demographic variable, and M is the number of variables included. These profiles are used to condition the LLM at the **prompt level**, guiding its initial understanding of the agent's characteristics. The specific prompts and the demographic features used are detailed in Appendix B.

3.1.2 Demographic Vectors. To enable stable and structured identity representation, we select representative demographic features that are broadly available across major social surveys, have stable definitions, and capture fundamental social positions, and then construct a demographic vector for each feature value. The specific features used are detailed in Appendix C.1.3. Let $\mathcal{V}_k$ denote the set of possible values for demographic variable k, and $v_{k,j} \in \mathcal{V}_k$ a specific value. The construction process is as follows:

Survey Question Simulation: For each demographic variable k, we first define a fixed set of survey questions $\{Q_k^{(1)}, \ldots, Q_k^{(R)}\}$ that probe the semantic implications of this attribute. For each value code $v_{k,j} \in V_k$, we then construct a set of value-specific prompts to elicit the model's internal representation of this identity. By combining the shared question set with these value-specific instructions, we generate a collection of synthetic prompts:

$$\{(Q_k^{(r)}, v_{k,j}^{(m)}) \mid r = 1, \ldots, R;\ m = 1, \ldots, M_{k,j}\},$$

where $v_{k,j}^{(m)}$ denotes the m-th role instruction associated with value $v_{k,j}$, and $M_{k,j}$ is the number of role instructions defined for that value. This results in $R \times \sum_j M_{k,j}$ question–response instances

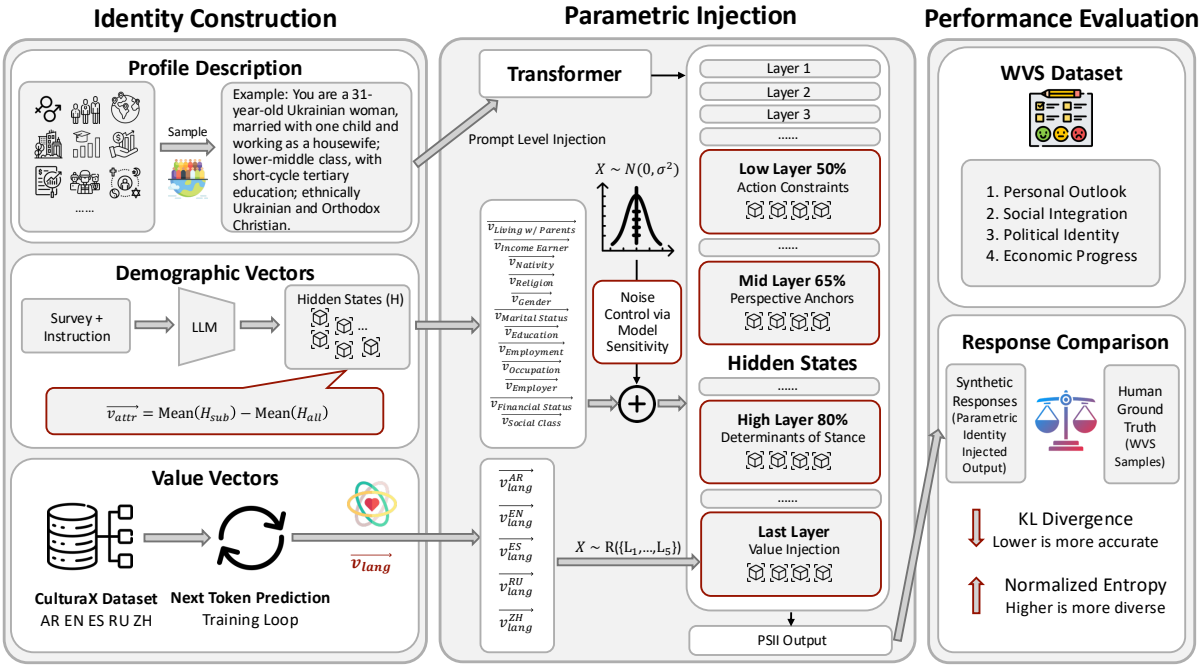


Figure 2: Overview of the Parametric Social Identity Injection (PSII) mechanism. From left to right: Identity Construction, including agent profile construction and identity vector construction; Parametric Injection, including noise addition and hierarchical injection; Performance Evaluation of the simulated agents.

for each demographic variable k, forming the demographic vector dataset, which is detailed in Appendix C.2.

LLM Response Embedding: For each synthetic prompt generated for demographic variable k and value $v_{k,j}$, the LLM produces a response. Let $h^{(l)}_{k,j,m,t}$ denote the hidden state of token t at layer l for the response generated from the m-th role instruction associated with value $v_{k,j}$. We compute the **layer-wise average hidden state** for each response as:

$$\bar{h}^{(l)}_{k,j,m} = \frac{1}{T_{k,j,m}} \sum_{t=1}^{T_{k,j,m}} h^{(l)}_{k,j,m,t},$$

where $T_{k,j,m}$ is the number of tokens in the corresponding response.

Value-Specific Vector Computation: The demographic vector $d_{k,j}$ is calculated as the difference between the mean response representation over all prompts conditioned on $v_{k,j}$ and the marginal mean over all values of the same demographic variable k:

$$\mathbf{d}_{k,j} = \frac{1}{|\mathcal{S}_{k,j}|} \sum_{(k,j,m)\in\mathcal{S}_{k,j}} \bar{h}^{(\mathcal{L}_k)}_{k,j,m} - \frac{1}{|\mathcal{S}_k|} \sum_{(k,j',m')\in\mathcal{S}_k} \bar{h}^{(\mathcal{L}_k)}_{k,j',m'},$$

where $\mathcal{S}_{k,j}$ denotes the set of all response instances generated from prompts conditioned on value $v_{k,j}$, $\mathcal{S}_k = \bigcup_j \mathcal{S}_{k,j}$ represents the union of these sets across all values of demographic variable k, and $\mathcal{L}_k$ is the layer selected for identity injection.

Additional analyses show that demographic-vector construction is robust across instruction models, prompt variants, and random seeds, and that the generated demographic semantics remain highly consistent across settings; see Appendix D.1.

3.1.3 Value Vectors. Language not only encodes information but also reflects cultural and worldview-specific patterns.

To approximate **value orientations** in the absence of explicit value annotations, we construct **language-based value vectors** using data from the target populations. Specifically, we learn a lightweight trainable vector $\mathbf{l}_s$ for each representative language s. Each value vector has the same dimensionality as the model hidden states, while all parameters of the base LLM are frozen. During training, $\mathbf{l}_s$ is optimized with the standard next-token prediction objective: adding $\mathbf{l}_s$ to the final-layer hidden state of the last token should increase the predictive likelihood of the next token in the corresponding language corpus. In this way, the learned vector captures language-specific distributional patterns and cultural-linguistic regularities, thereby anchoring generated responses within culturally informed reasoning frameworks without updating the base model.

3.2 Parametric Injection

PSII injects identity information at two distinct levels and treats demographic and language vectors differently.

At the **prompt level**, agent profiles P_i are included in the input prompt to provide semantic context. This ensures that the model has initial knowledge of the agent's demographic attributes before generation begins.

At the **representation level**, identity vectors are injected directly into the hidden states of the LLM. When predicting a response for a given sample, we first extract the demographic values j_k for all variables k and select the corresponding demographic vectors $\mathbf{d}_{k,j_k}$. During forward propagation, these vectors are injected into

the hidden states of specific layers $\mathcal{L}_k$ using forward hooks. For token t at layer $\mathcal{L}_k$, the hidden state is updated as:

$$\tilde{h}_t^{(\mathcal{L}_k)} = h_t^{(\mathcal{L}_k)} + \mathbf{d}_{k,j_k} + \boldsymbol{\epsilon}_t,$$

where $\boldsymbol{\epsilon}_t \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ is a small Gaussian noise vector used to induce intra-group heterogeneity. In practice, demographic vectors are constructed separately for each Transformer layer. At injection time, we select the vector corresponding to the target injection layer, so no vector is shared across layers. During prompt encoding, the demographic vector is added to all prompt-token representations to establish a global demographic condition. During autoregressive generation, the same vector is added to the hidden representation of the newly generated token at each step, thereby continuously steering the response.

To incorporate diverse cultural and linguistic context, we randomly select a language s for each sample, translate the prompt into that language, and inject the corresponding language vector $\mathbf{l}_s$ at the **last layer** L. For token t at the last layer during generation, the hidden state is updated as:

$$\tilde{h}_t^{(L)} = h_t^{(L)} + \mathbf{l}_s.$$

This mechanism introduces language-specific expression patterns and reasoning tendencies learned from the corpus, effectively providing a culturally informed anchor for model outputs.

3.2.1 Noise Module. To model intra-group heterogeneity and prevent the over-essentialization of identity, we introduce a controlled Gaussian noise vector $\boldsymbol{\epsilon}_t$ when injecting demographic vectors. This noise adds small random perturbations to each demographic vector, simulating individual differences among agents with the same demographic attributes, thereby partially addressing the problem of insufficient internal diversity.

$$\boldsymbol{\epsilon}_t \sim \mathcal{N}(0, \sigma^2\mathbf{I}),$$

where σ is the standard deviation of the Gaussian noise, controlling the magnitude of perturbation applied to demographic vectors. Ideally, σ should be calibrated to introduce variability without disrupting the model's core reasoning capabilities or demographic consistency. The specific calibration strategy and parameter selection are detailed in Section 4.4.

3.2.2 Layer-wise Hierarchical Injection. Transformer-based LLMs capture information at different levels across their internal representations, ranging from linguistic constraints to abstract reasoning. Empirical studies in cognitive modeling and LLM behavior indicate that lower layers typically encode surface-level patterns, as well as syntactic and stylistic features; intermediate layers capture contextual information and background assumptions; and upper layers are responsible for abstract reasoning, value judgments, and final decision-making. Moreover, different personality traits and demographic attributes are processed differently across Transformer layers. Motivated by these observations, we adopt a hierarchical injection strategy, inserting demographic vectors into the layers that most naturally align with their semantic processing roles. Based on empirical analysis (see Section 5.4 and Appendix E.3 for details), demographic attributes are categorized into three hierarchical layer groups:

Lower layers: Govern behavioral feasibility and constraint processing. Demographic information related to responsibilities, life structure, or family obligations is injected at this level. Examples include: *living with parents, primary income earner.*

Intermediate layers: Determine perspective and guide problem framing. Attributes reflecting experience sources, normative assumptions, or background context are injected at this level. Examples include: *religion, immigration status.*

Upper layers: Govern final stance, value judgments, and decision level outputs. Attributes defining status, ideology, or structured social identity are injected at this level. Examples include: *gender, marital status, education, employment, occupation, employer type, financial status, social class.*

3.3 Theoretical Rationale

The effectiveness of Parametric Social Identity Injection in maintaining stable demographic identities can be understood through the lens of representation-level control and the shortcomings of traditional persona modeling. Existing approaches typically rely on in-context learning (ICL) via textual prompts to induce individual differences. While prompt-based conditioning can guide the model at a surface level, it suffers from identity decay during long text generation or complex reasoning: as hidden representations propagate through multiple layers, the semantic impact of the prompt diminishes, leading to homogenized outputs and reduced diversity.

PSII mitigates this limitation by explicitly injecting demographic vectors into intermediate hidden states. These vectors act as structured, persistent constraints that continuously modulate token-level activations across layers, ensuring that each generated response adheres to the demographic identity of the synthetic agent. This mechanism fosters logical consistency and demographic stability, producing agents whose behavior is coherent, heterogeneous, and structured rather than merely repeating superficial prompt cues.

Additionally, the diversity of the injected vectors introduces controlled intra-group variability. By incorporating small, stochastic perturbations, PSII simulates natural heterogeneity within demographic categories, addressing the problem of identity essentialism and further enhancing the realism of the generated population. In effect, PSII provides both a stable anchor for identity and a flexible mechanism for capturing nuanced inter- and intra-group variation.

4 Experimental Setup

4.1 Dataset and Evaluation Metrics

We conduct experiments on the **World Values Survey (WVS)** dataset, a large-scale, cross-national survey designed to measure human values, beliefs, and socio-political attitudes across diverse cultural and demographic groups (See Appendix C.1.1 for details). Following prior work, we use responses to **Q1–Q259** as target opinion questions, while the remaining questions **Q260–Q290** are reserved for demographic feature modeling and identity alignment.

For analysis, the 259 opinion questions are grouped into four high-level categories based on their thematic content: Personal Beliefs & Life Outlook, Social Integration & Perception, Political Engagement & Institutional Identity, and Economic Development & Progress. This regrouping is theory-driven and informed by

established value-dimension frameworks, including the Inglehart–Welzel cultural map, as well as prior WVS-based simulation studies. Details of the regrouping rationale are provided in Appendix C.1.2.

Model performance is evaluated at the group level by comparing the distribution of model-generated responses with the human ground-truth distribution. We report **KL divergence** to measure distributional accuracy, where lower values indicate better alignment with human responses.

To assess diversity, we compute **Entropy Deviation (ED)**, defined as the absolute difference between the normalized entropy of model-generated responses and that of human responses, where a lower ED indicates closer diversity matching to the human distribution.

4.2 Baseline Methods

We compare our method against a diverse set of representative baselines:

- **Direct**: Direct simulation without any fine-tuning or prompt engineering [31].
- **High-Temp**: High-temperature sampling to encourage output variability, with temperature = 2 [9].
- **Multilingual**: Multilingual prompting to induce diversity by varying the language context [39].
- **DivReq**: Explicitly requesting diversity in the prompt, without additional structural constraints [39].
- **PE**: Prompt engineering with carefully designed persona templates to better approximate human respondents [42].
- **SimVBG**: The SimVBG method [11], which constructs background stories and is guided by the Cognitive-Affective Personality System (CAPS) theory.
- **PV**: Persona Vectors [7], a representation-level steering method that identifies persona-related directions in the model's activation space and uses them to control generated character traits.

Detailed implementations of all baseline methods are described in Appendix B.

4.3 Implementation Details

We evaluate all methods on four instruction-tuned large language models: **Qwen2.5-7B-Instruct**, **Qwen2.5-14B-Instruct**, **Llama-3.1-8B-Instruct**, and **Mistral-24B-Instruct**.

From the full WVS dataset containing 97,220 respondents, we randomly sample 100 individuals to construct simulated agent populations for evaluation. Additional resampling experiments confirm that PSII is robust to random sampling variation, as shown in Appendix D.2. Each simulated agent answers one question at a time, following the original WVS question order, and produces a numerical response consistent with the survey format.

To approximate value orientations, we train language-specific value vectors using the CulturaX dataset [27] for the five primary languages $s \in \{en, zh, ar, es, ru\}$ (See Appendix C.1.4), with training hyperparameters set to $n\_samples = 20000$, $epochs = 3$, and $learning\_rate = 1 \times 10^{-3}$.

Unless otherwise specified, generation uses default decoding parameters with temperature = 0.7 and top_k = 20.

For methods involving stochastic identity perturbation, we inject Gaussian noise into the demographic vectors, with model-specific standard deviations: Qwen2.5-14B ($\sigma = 0.35$), Qwen2.5-7B ($\sigma = 0.30$), Mistral-24B ($\sigma = 0.09$), and Llama-3.1-8B ($\sigma = 0.07$).

Training of identity vectors is performed on a single **NVIDIA A100-SXM4-80GB** GPU, while inference can be completed on a single **NVIDIA A100-SXM4-40GB** GPU.

4.4 Noise Calibration Strategy

To determine the optimal noise standard deviation σ for each model, we propose a calibration metric termed **model sensitivity**. This metric quantifies the impact of noise on the predicted ranking of response options.

Specifically, for each agent i and question j, we compute the Mean Absolute Error (MAE) between the rankings of answer options predicted with and without noise, where the ranking reflects the position of each option among all candidate options for that specific question:

$$MAE_{i,j} = \frac{\left|\text{rank}(\text{answer}_{i,j}^{\text{noise}}) - \text{rank}(\text{answer}_{i,j}^{\text{no noise}})\right|}{\text{number of options}}.$$

The overall model sensitivity is computed by averaging over all sampled agents and questions. A higher sensitivity value indicates that the model's reasoning is easily disrupted by perturbations, requiring a smaller σ.

Empirically, we observed a linear correlation between the optimal noise level and model robustness. We specifically calibrate the noise standard deviation as:

$$\sigma_{\text{best}} = \max(0, 0.4 - \text{model sensitivity}).$$

Based on this calibration, if the model is too sensitive (sensitivity > 0.4), we set $\sigma = 0$. The specific calibrated σ values for each model used in our experiments are reported in Section 4.3.

5 Experiments and Results

5.1 Main Results

We evaluate different simulation methods by measuring how well they reproduce human survey outcomes. The main results are summarized in Table 1. Overall, PSII consistently achieves the best trade-off between accuracy and diversity across models, parameter settings, and question categories, substantially narrowing the gap between simulated and human survey responses. See Appendix E.1 for additional results.

Under direct prompting, Mistral-24B produces responses most similar to human data, followed by Llama-3.1-8B and then the Qwen series. After incorporating PSII, all models exhibit marked gains in both accuracy and diversity. Among them, Llama-3.1-8B benefits the most from PSII, followed by Qwen2.5-7B.

Performance also varies across question subsets. PSII performs best on the Economic Progress questions, while performance on Beliefs & Life questions is relatively weaker. This suggests that questions related to economic evaluations may align more naturally with the structured representation shifts induced by PSII, while questions related to beliefs and values remain more challenging to simulate faithfully.

Table 1: Main experimental results on the WVS dataset. We report KL divergence and Entropy Deviation (ED) for each method across four question categories and overall. Best-performing results are highlighted in bold.

Model	Method	Beliefs & Life		Social Integration		Political Engagement		Economic Progress		Overall	
		KL ↓	ED ↓	KL ↓	ED ↓	KL ↓	ED ↓	KL ↓	ED ↓	KL ↓	ED ↓
Qwen2.5-7B	Direct	1.5182	0.6664	0.8575	0.7406	2.2222	0.7625	1.6283	0.7932	1.3915	0.7340
	High-Temp	1.1941	0.4754	0.7074	0.5981	1.9653	0.6266	1.1072	0.5272	1.1605	0.5778
	Multilingual	1.0946	0.3706	0.5547	0.5112	1.9297	0.5000	1.1030	0.5371	1.0568	0.4811
	DivReq	1.5442	0.6671	0.8383	0.7464	2.2166	0.7587	1.5282	0.7283	1.3813	0.7329
	PE	1.4812	0.4050	0.7182	0.6292	1.8852	0.4847	1.5095	0.6046	1.2209	0.5443
	SimVBG	0.8420	0.2494	0.3812	0.2942	1.0759	0.3025	1.0954	0.3646	0.6945	0.2908
	PV	1.5973	0.4796	0.7220	0.6551	1.8013	0.6563	0.9621	0.4417	1.1982	0.6102
	PSII	**0.6772**	**0.0989**	**0.1862**	**0.0180**	**0.9095**	**0.0150**	**0.3094**	**0.0150**	**0.4843**	**0.0319**
Qwen2.5-14B	Direct	1.1020	0.7509	0.8343	0.7654	2.5925	0.7812	1.6882	0.8866	1.3982	0.7724
	High-Temp	0.9247	0.6179	0.7293	0.6685	2.4167	0.7003	1.2345	0.6439	1.2435	0.6657
	Multilingual	0.8584	0.5626	0.5423	0.5110	2.0181	0.5082	1.0812	0.6043	1.0258	0.5251
	DivReq	1.2082	0.7718	0.8749	0.7751	2.6916	0.7556	1.7217	0.8545	1.4673	0.7730
	PE	0.7451	0.4255	0.6418	0.5580	1.4125	0.4700	1.1097	0.4664	0.8905	0.5035
	SimVBG	**0.4102**	0.2537	0.3562	0.2736	1.6102	0.3719	0.6910	0.2851	0.7215	0.2967
	PV	0.9186	0.6031	0.6655	0.6207	2.7785	0.6194	1.0715	0.5878	1.3005	0.6153
	PSII	0.5193	**0.1264**	**0.3247**	**0.2540**	**1.1101**	**0.2112**	**0.4835**	**0.1540**	**0.5814**	**0.2123**
Llama-3.1-8B	Direct	1.1033	0.6411	0.8232	0.6060	2.2543	0.6592	1.3212	0.7199	1.2856	0.6326
	High-Temp	0.7738	0.3037	0.5194	0.3257	1.7162	0.3421	0.6542	0.3198	0.8970	0.3254
	Multilingual	0.7657	0.4619	0.5298	0.3693	1.6478	0.3872	1.1913	0.6584	0.9071	0.4063
	DivReq	1.0864	0.5541	0.9328	0.6202	1.8263	0.5825	1.2447	0.6700	1.2172	0.5992
	PE	0.7124	0.3630	0.5219	0.4198	1.4478	0.3623	0.5560	0.2065	0.8095	0.3831
	SimVBG	0.4275	0.2284	0.4879	0.2240	1.0170	0.3074	0.7253	0.2610	0.6298	0.2491
	PV	1.0344	0.5709	0.6095	0.4850	1.7505	0.5276	1.1692	0.6609	1.0263	0.5220
	PSII	**0.3776**	**0.0157**	**0.2305**	**0.0254**	**0.7494**	**0.0386**	**0.2911**	**0.0601**	**0.4017**	**0.0040**
Mistral-24B	Direct	0.8621	0.4228	0.7749	0.5967	1.5736	0.5651	0.9845	0.4081	1.0158	0.5445
	High-Temp	0.6893	0.1083	0.3788	0.0940	1.2904	0.2048	0.8349	0.2820	0.7064	0.1353
	Multilingual	0.6373	0.2698	0.4888	0.3609	1.4458	0.3030	0.6964	0.3057	0.7843	0.3246
	DivReq	0.8306	0.1917	0.6905	0.3010	1.1870	0.3541	1.0052	0.2957	0.8662	0.2930
	PE	0.6449	0.2925	0.6518	0.5587	1.3570	0.4041	1.0166	0.3885	0.8560	0.4558
	SimVBG	**0.3285**	0.2228	0.3066	0.2458	1.5274	0.2364	0.6564	0.1738	0.6571	0.2353
	PV	0.7413	0.3592	0.6293	0.5850	1.7421	0.4808	0.7575	0.3323	0.9555	0.4999
	PSII	0.4795	**0.0013**	**0.2278**	**0.0880**	**1.2505**	**0.1244**	**0.4134**	**0.0252**	**0.5607**	**0.0774**

Compared to baseline methods, PSII reliably outperforms simple strategies such as Direct and Diversity Request. High-temperature decoding performs relatively well on Mistral-24B, but degrades substantially on other models, indicating that uncontrolled randomness alone does not provide a robust or general solution for realistic opinion simulation. Relative to Multilingual prompting and PE, PSII further introduces representation-level identity vector injection, resulting in better simulations. PSII also improves upon Persona Vectors by further incorporating value vectors, prompt-based profiles, controlled noise, and hierarchical layer-wise injection. Although SimVBG is a strong background-story-based baseline, PSII consistently achieves better overall performance across most settings.

Finally, we observe a noticeable gap in ED achieved by PSII between Qwen2.5-7B and Qwen2.5-14B, which we attribute primarily to scale-dependent differences in internal representations. As PSII directly intervenes in hidden states, its impact on diversity is sensitive to model depth, hidden dimensionality, and representation geometry, which differ substantially across model scales. In practice, different scales also exhibit varying sensitivity to the injected noise, and the noise variance is therefore selected separately for each model to ensure stable generation, which may further contribute to the ED variation. Moreover, although the injection depth is aligned by relative position, differences in absolute depth imply that the injected layers may correspond to different functional stages in the two models, leading to specific diversity effects.

5.2 Response Distribution Analysis

To further analyze how different simulation methods capture behavioral variability, we examine the distributions of individual response choices, aiming to assess whether the response patterns

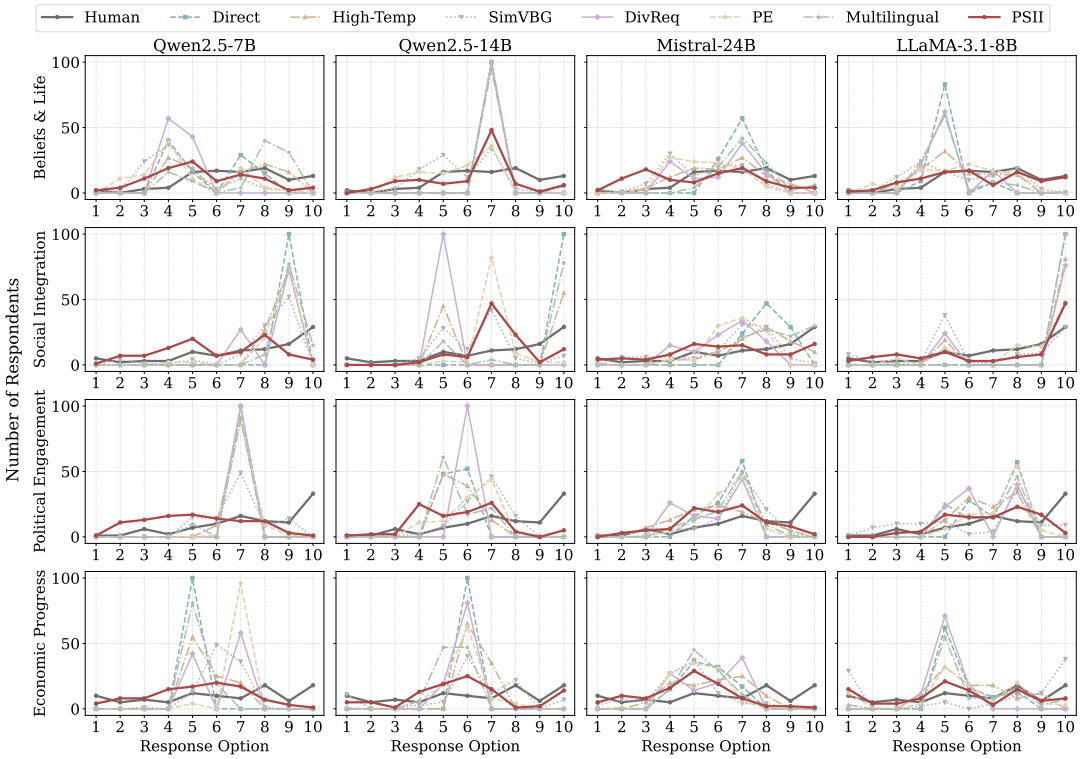


Figure 3: Response distributions for a randomly selected question from each of the four categories. PSII more closely matches the empirical response diversity observed in human survey data, while baseline methods often concentrate on a few options.

Table 2: Ablation study results on PSII across different models. Each row shows the impact of removing one component on accuracy and diversity metrics.

Setting	Qwen2.5-7B		Qwen2.5-14B		Llama-3.1-8B		Mistral-24B	
	KL ↓	ED ↓	KL ↓	ED ↓	KL ↓	ED ↓	KL ↓	ED ↓
PSII (full model)	**0.4843**	0.0319	**0.5814**	**0.2123**	**0.4017**	**0.0040**	**0.5607**	0.0774
PSII w/o value vector	0.5264	0.0348	0.6574	0.3040	0.4638	0.0416	0.6254	0.1723
PSII w/o demographic vectors	0.9705	0.3859	0.7893	0.3878	0.6681	0.2729	0.7185	0.3312
PSII w/o prompt-based profile	0.5024	**0.0151**	0.6688	0.2527	0.5575	0.0630	0.6149	**0.0474**
PSII w/o parametric noise	0.7237	0.2261	0.7332	0.3294	0.5673	0.2068	0.6813	0.2808
PSII w/o layer-wise injection	0.5313	0.0964	0.7665	0.2849	0.5225	0.1089	0.5780	0.0999

generated by simulated agents reflect the dispersion observed in human survey data.

Specifically, we randomly select one question from each of the four question categories and visualize the response distributions produced by 100 simulated agents under different methods, alongside the corresponding empirical distributions from human respondents.

As shown in Figure 3, baseline methods tend to concentrate responses on a small number of options, exhibiting limited diversity. In contrast, PSII produces more evenly distributed responses that more closely match human data, indicating that it better preserves inter-agent heterogeneity at the level of concrete survey responses. These findings are consistent with our quantitative results and further validate the effectiveness of PSII in enhancing diversity.

5.3 Ablation Studies

To investigate the contribution of each component in PSII, we perform ablation studies in which we remove one module at a time, including the value vector, demographic vectors, prompt-based profile, parametric noise, and the layer-wise injection (modified to inject at 70% of layers). The results are summarized in Table 2 and report both simulation accuracy and diversity metrics across different models.

The results show that removing any single component reduces simulation accuracy, and most ablations also degrade diversity

matching, indicating that all modules contribute meaningfully to PSII's overall effectiveness. Among them, the demographic vectors are the most critical for maintaining both accuracy and diversity, while parametric noise is especially important for enhancing diversity. Both the value vector and the layer-wise injection significantly contribute to gains in both accuracy and diversity. Prompt-based profiles act as semantic anchors, improving alignment to the target distribution but may constrain the output space, thus reducing diversity. This reflects a trade-off between accuracy and diversity. PSII combines prompt- and representation-level steering. While removing this module can possibly increase diversity, our ultimate goal is accurate distribution matching rather than maximizing diversity; therefore, retaining this module is the preferred design choice.

Although the sensitivity to each component varies slightly across models, the overall trends are consistent, demonstrating that PSII is robust and effective across different LLM backbones.

These findings highlight that PSII's performance arises from the complementary effects of its modules: structured identity and value representations provide coherent guidance for realistic responses, while carefully injected noise ensures sufficient behavioral variability, and layered interventions allow these effects to propagate effectively through the model.

5.4 Layer-wise Injection Analysis

Before applying hierarchical injection in our full PSII framework, we conducted experiments on the Qwen2.5-7B model to determine the most suitable layers for injecting different personality attributes. Specifically, we tested all demographic attributes injected individually across layers 1 to 28, and recorded their impact on simulation performance, as shown in Appendix E.2, Figure 5.

As illustrated, the optimal injection layer varies across different attributes. We selected the best layer for each attribute based on the layer that minimized the KL divergence, indicating the closest alignment with human response distributions. This approach ensures that each demographic attribute is incorporated at a point in the network where it most effectively influences the model's output without disrupting stability.

Furthermore, we observed that when all vectors are injected simultaneously at a fixed layer, injecting at approximately 70% of the network depth achieves the best overall performance. This finding motivated our choice in the ablation studies, where we conducted comparisons using injection at this 70% layer as a representative baseline.

6 Ethical Concerns and Societal Implications

While PSII enables scalable and diversity-aware social simulation, its deployment requires careful consideration of several ethical and societal implications.

A key concern involves data usage and privacy. Our experiments rely exclusively on publicly available human value datasets (e.g., World Values Survey), ensuring that no private or sensitive individual data is accessed or inferred. Researchers must continue to prioritize privacy-preserving practices when extending PSII to other datasets.

Another risk lies in demographic modeling, which may introduce issues such as stereotyping or oversimplification. To mitigate these, we: (a) strictly follow established survey definitions; (b) focus on group-level (not individual-level) analysis; and (c) apply manual filtering. PSII is intended for controlled research settings only, not for real-world decision-making.

Beyond modeling choices, the responsible use of synthetic agents is critical. Although LLM-based agents can simulate large-scale public opinion efficiently, unregulated deployment may lead to misuse, such as generating synthetic narratives for propaganda or social manipulation. Institutional and technical safeguards are therefore essential to prevent abuse, including access control, transparency reports, and usage auditing.

Finally, context-aware application must be ensured. PSII-generated simulations should be interpreted as supplementary models rather than definitive reflections of societal opinions. Users are expected to consider limitations such as cultural nuances, minority representation, and model bias, and to carefully design experiments that mitigate potential harm or misinterpretation.

7 Conclusion

This study investigates a key challenge faced by large language models in public opinion simulation, specifically the phenomenon of "Diversity Collapse," where synthetic populations exhibit significant inter-group homogenization and insufficient intra-group representativeness. By analyzing the internal representation dynamics, we reveal that this issue stems, in part, from a systematic contraction of representations in the upper layers of the Transformer, causing distinct social identities to converge toward a uniform state at the end of the reasoning chain. To address this, we propose the Parametric Social Identity Injection (PSII) framework. By directly embedding parametric vectors of demographic attributes and value orientations into the intermediate hidden states, PSII achieves stable guidance and fine-grained modulation of identity attributes at the representation level. This allows identity signals to persist throughout generation and enables structured, controllable diversity aligned with population attributes. Extensive experiments on the World Values Survey (WVS) demonstrate that PSII consistently enhances the distributional fidelity and diversity of simulation results across multiple mainstream open-source LLMs, significantly outperforming existing baseline methods. Our work not only elucidates the impact of model representations on simulation quality but also provides a practical technical pathway for constructing large-scale, high-fidelity, and diversity-aware digital twin social simulations.

Acknowledgments

This work is supported by the Research Project of Quancheng Laboratory, China (Grant No. QCL20250105), the National Natural Science Foundation of China No. 62502260, and the Postdoctoral Fellowship Program of China Postdoctoral Science Foundation No. GZC20240833.

References

[1] Axel Abels and Tom Lenaerts. 2025. Wisdom from Diversity: Bias Mitigation Through Hybrid Human-LLM Crowds. *arXiv preprint arXiv:2505.12349* (2025).

[2] 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 (2023), 337–351.

[3] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609* (2023).

[4] Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2024. Sensitivity, Performance, Robustness: Deconstructing the Effect of Sociodemographic Prompting. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, Yvette Graham and Matthew Purver (Eds.). Association for Computational Linguistics, St. Julian's, Malta, 2589–2615. doi:10.18653/v1/2024.eacl-long.159

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

[6] Julien Boelaert, Samuel Coavoux, Etienne Ollion, Ivaylo Petev, and Patrick Präg. 2025. Machine Bias. How Do Generative Language Models Answer Opinion Polls? *Sociological Methods & Research* (2025), 00491241251330582.

[7] Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. 2025. Persona vectors: Monitoring and controlling character traits in language models. *arXiv preprint arXiv:2507.21509* (2025).

[8] Rongxin Chen, Tianyu Wu, Bingbing Xu, Xiucheng Xu, and Huawei Shen. 2026. HAG: Hierarchical Demographic Tree-based Agent Generation for Topic-Adaptive Simulation. *arXiv preprint arXiv:2601.05656* (2026).

[9] John Chung, Ece Kamar, and Saleema Amershi. 2023. Increasing diversity while maintaining accuracy: Text data generation with large language models and human interventions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 575–593.

[10] Don A Dillman, Jolene D Smyth, and Leah Melani Christian. 2014. Internet, phone, mail, and mixed-mode surveys: The tailored design method. *Indianapolis, Indiana* 17 (2014).

[11] Bangde Du, Ziyi Ye, Zhijing Wu, Monika A. Jankowska, Shuqi Zhu, Qingyao Ai, Yujia Zhou, and Yiqun Liu. 2025. SimVBG: Simulating Individual Values by Backstory Generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (Eds.). Association for Computational Linguistics, Suzhou, China, 13093–13122. doi:10.18653/v1/2025.emnlp-main.662

[12] Alessandro Fabris, Stefano Messina, Gianmaria Silvello, and Gian Antonio Susto. 2022. Algorithmic fairness datasets: the story so far. *Data Mining and Knowledge Discovery* 36, 6 (2022), 2074–2152.

[13] Robert M Groves, Floyd J Fowler Jr, Mick P Couper, James M Lepkowski, Eleanor Singer, and Roger Tourangeau. 2011. *Survey methodology*. John Wiley & Sons.

[14] Robert M Groves and Lars Lyberg. 2010. Total survey error: Past, present, and future. *Public opinion quarterly* 74, 5 (2010), 849–879.

[15] Shirley Anugrah Hayati, Minhwa Lee, Dheeraj Rajagopal, and Dongyeop Kang. 2024. How far can we extract diverse perspectives from large language models?. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 5336–5366.

[16] Babak Hemmatian and Lav R Varshney. 2022. Debiased large language models still associate muslims with uniquely violent acts. *arXiv preprint arXiv:2208.04417* (2022).

[17] Zhengyu Hu, Jianxun Lian, Zheyuan Xiao, Max Xiong, Yuxuan Lei, Tianfu Wang, Kaize Ding, Ziang Xiao, Nicholas Jing Yuan, and Xing Xie. 2025. Population-aligned persona generation for llm-based social simulation. *arXiv preprint arXiv:2509.10127* (2025).

[18] Ji Huang, Mengfei Li, and Shuai Shao. 2025. Distribution Shift Alignment Helps LLMs Simulate Survey Response Distributions. *arXiv preprint arXiv:2510.21977* (2025).

[19] EunJeong Hwang, Bodhisattwa Majumder, and Niket Tandon. 2023. Aligning Language Models to User Opinions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 5906–5919. doi:10.18653/v1/2023.findings-emnlp.393

[20] Carolin Kaiser, Jakob Kaiser, Vladimir Manewitsch, Lea Rau, and Rene Schallner. 2025. Simulating Human Opinions with Large Language Models: Opportunities and Challenges for Personalized Survey Data Modeling. In *Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization*. 82–86.

[21] Rabimba Karanjai, Boris Shor, Amanda Austin, Ryan Kennedy, Yang Lu, Lei Xu, and Weidong Shi. 2025. Synthesizing Public Opinions with LLMs: Role Creation, Impacts, and the Future to eDemorcacy. *arXiv preprint arXiv:2504.00241* (2025).

[22] Ayato Kitadai, Kazuhito Ogawa, and Nariaki Nishino. 2024. Examining the feasibility of large language models as survey respondents. In *2024 IEEE International Conference on Big Data (BigData)*. IEEE, 3858–3864.

[23] Ivar Krumpal. 2013. Determinants of social desirability bias in sensitive surveys: a literature review. *Quality & quantity* 47, 4 (2013), 2025–2047.

[24] Preethi Lahoti, Nicholas Blumm, Xiao Ma, Raghavendra Kotikalapudi, Sahitya Potluri, Qijun Tan, Hansa Srinivasan, Ben Packer, Ahmad Beirami, Alex Beutel, et al. 2023. Improving diversity of demographic representation in large language models via collective-critiques and self-voting. *arXiv preprint arXiv:2310.16523* (2023).

[25] Mistral AI. 2025. Mistral Small 3. https://mistral.ai/news/mistral-small-3/.

[26] A Myers. 2021. Rooting out anti-Muslim bias in popular language model GPT-3. *Stanford HAI* (2021).

[27] Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. 2024. Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 4226–4237.

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

[29] John Platt et al. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers* 10, 3 (1999), 61–74.

[30] Yao Qu and Jue Wang. 2024. Performance and biases of large language models in public opinion simulation. *Humanities and Social Sciences Communications* 11, 1 (2024), 1–13.

[31] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect?. In *International Conference on Machine Learning*. PMLR, 29971–30004.

[32] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. AI models collapse when trained on recursively generated data. *Nature* 631, 8022 (2024), 755–759.

[33] Weihang Su, Yichen Tang, Qingyao Ai, Junxi Yan, Changyue Wang, Hongning Wang, Ziyi Ye, Yujia Zhou, and Yiqun Liu. 2025. Parametric Retrieval Augmented Generation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Padua, Italy) *(SIGIR '25)*. Association for Computing Machinery, New York, NY, USA, 1240–1250. doi:10.1145/3726302.3729957

[34] Joseph Suh, Erfan Jahanparast, Suhong Moon, Minwoo Kang, and Serina Chang. 2025. Language Model Fine-Tuning on Scaled Survey Data for Predicting Distributions of Public Opinions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 21147–21170. doi:10.18653/v1/2025.acl-long.1028

[35] Roger Tourangeau, Lance J Rips, and Kenneth Rasinski. 2000. The psychology of survey response. (2000).

[36] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[37] Angelina Wang, Jamie Morgenstern, and John P Dickerson. 2025. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence* (2025), 1–12.

[38] Mengxin Wang, Dennis J Zhang, and Heng Zhang. 2024. Large language models for market research: A data-augmentation approach. *arXiv preprint arXiv:2412.19363* (2024).

[39] Qihan Wang, Shidong Pan, Tal Linzen, and Emily Black. 2025. Multilingual Prompting for Improving LLM Generation Diversity. *arXiv preprint arXiv:2505.15229* (2025).

[40] Justin Wong, Yury Orlovskiy, Michael Luo, Sanjit A Seshia, and Joseph E Gonzalez. 2024. Simplestrat: Diversifying language model generation with stratification. *arXiv preprint arXiv:2410.09038* (2024).

[41] Kaiqi Yang, Hang Li, Hongzhi Wen, Tai-Quan Peng, Jiliang Tang, and Hui Liu. 2024. Are Large Language Models (LLMs) Good Social Predictors?. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 2718–2730. doi:10.18653/v1/2024.findings-emnlp.153

[42] Kaiqi Yang, Hang Li, Hongzhi Wen, Tai-Quan Peng, Jiliang Tang, and Hui Liu. 2024. Are Large Language Models (LLMs) Good Social Predictors? *arXiv preprint arXiv:2402.12620* (2024).

[43] Lily Hong Zhang, Smitha Milli, Karen Jusko, Jonathan Smith, Brandon Amos, Wassim Bouaziz, Manon Revel, Jack Kussman, Yasha Sheynin, Lisa Titus, et al. 2025. Cultivating pluralism in algorithmic monoculture: The community alignment dataset. *arXiv preprint arXiv:2507.09650* (2025).

[44] Yujia Zhou, Hexi Wang, Qingyao Ai, Zhen Wu, and Yiqun Liu. [n. d.]. Investigating Prosocial Behavior Theory in LLM Agents under Policy-Induced Inequities. *arXiv preprint arXiv* 2505 ([n. d.]).

A Limitations

Despite its advantages, PSII has several limitations that warrant attention.

Dependence on demographic coverage. The quality of synthetic populations is constrained by the granularity and completeness of available demographic data. Rare or underrepresented groups may still be insufficiently modeled, potentially limiting simulation fidelity in these populations.

Scope of identity modeling. Current PSII vectors primarily encode coarse-grained demographic attributes and value orientations. Fine-grained personality traits, dynamic opinion changes, or context-specific behavioral nuances are not explicitly captured and may require additional mechanisms.

B Baseline Implementation Details

We select baseline methods from prior work that are applicable to the World Values Survey (WVS) setting. Specifically, we focus on approaches designed for population-level simulation and structured or semi-structured survey settings. Several diversity-enhancing methods proposed in the literature primarily target semantic or stylistic diversity in open-ended text generation, which are not suitable for closed-form, fixed-choice survey questions such as those in WVS. These methods are therefore excluded from our comparison.

Below, we describe the implementation details of each baseline considered in our experiments.

Direct. The Direct baseline performs LLM-based simulation without any explicit mechanisms for diversity control or identity conditioning. For each survey question, the model is prompted to generate a response directly, serving as a minimal and commonly used reference setting in prior social simulation work.

Direct Prompt

Forget you are an AI model. Simulate a human being.

High-Temp. The High-Temp baseline applies high-temperature sampling to encourage output variability. It uses the same prompting strategy as Direct, but sets the sampling temperature to 2. This baseline tests whether stochastic decoding alone is sufficient to induce population-level diversity.

Multilingual. The Multilingual baseline aims to increase diversity through linguistic variation. The original prompt used in Direct is translated into five languages: Arabic (ar), English (en), Spanish (es), Russian (ru), and Chinese (zh). For each simulated individual, one language is randomly selected using a fixed random seed (42). Apart from the prompt language, all other settings are identical to the Direct baseline.

DivReq. The DivReq baseline introduces an explicit diversity request at the prompt level. Specifically, we augment the Direct prompt with an additional instruction encouraging output diversity. This baseline assesses whether prompt-level diversity requests alone are sufficient to alleviate output homogenization.

Additional Instruction

Please try to be as diverse as possible.

PE (Prompt Engineering). The PE baseline conditions the model on structured demographic profiles constructed from real human samples in the WVS dataset. Specifically, we convert the self-reported demographic information of each subject in the dataset (e.g., age, gender, income level, education, religious beliefs, etc.) into natural language descriptions, which are then used as prompts to input into a LLM, thereby guiding the model to simulate the responses, attitudes, or behavioral reactions of that individual.

PE Prompt

Please answer based on the following personal profile:
You are a {sex}. You are aged {age}. You live in {country}. You are {citizenship}. {immigrant_self_status}. {immigrant_mother_status}. {immigrant_father_status}. You live in a {urban_rural} area, {settlement_type}, with a population of {settlement_size}. Your household consists of {household_size} members. You {live_with_parents}. You speak {home_language} at home. You are {marital_status}, have {num_children} children. Your education level is {education_self}. Your spouse's education level is {education_spouse}. Your father's education level is {education_father}. Your mother's education level is {education_mother}. You are {employment_self}. You {occupation_self}. You work in the {work_sector}. Your spouse is {employment_spouse}. Your spouse {occupation_spouse}. Your father {occupation_father14} when you were 14 years old. You {chief_wage_earner}. During the past year, your family {family_savings}. You belong to the {social_class}. Your income is in income group {income_decile} (1 = lowest, 10 = highest). {religion_status}. Your ethnic group is {ethnic_group}.

PE Example

Please answer based on the following personal profile:
You are a female. You are aged 35. You live in Chile. You are a citizen of this country. You are born locally. Your mother is born locally. Your father is born locally. You live in a urban area, regional center, with a population of 100,000-500,000. Your household consists of 2 members. You do not live with parents. You speak Spanish; Castilian at home. You are living together as married, have 0 children. Your education level is Master's degree. Your spouse's education level is Master's degree. Your father's education level is upper secondary education. Your mother's education level is upper secondary education. You are employed full-time. You work in professional/technical fields. You work in the private business/industry. Your spouse is employed full-time. Your spouse works in professional/technical fields. Your father worked as a semi-skilled worker when you were 14 years old. You are not the chief wage earner. During the past year, your family had spent some savings. You belong to the lower middle class. Your income is in income group group 5 (1 = lowest, 10 = highest). You identify as Roman Catholic. Your ethnic group is White, Caucasian.

This baseline represents a strong prompt-based identity conditioning approach commonly used in prior social simulation work. Similarly, in PSII, the prompt-level injection of demographic information is implemented in the same manner.

SimVBG. SimVBG first converts structured demographic profiles into a coherent background narrative. Guided by the Cognitive-Affective Personality System (CAPS) theory, it then generates candidate responses independently along three dimensions: cognitive, affective, and behavioral, and aggregates them to produce the final

simulated response. We follow the original paper's implementation and parameter settings [11].

PV. Persona Vectors [7], which steer LLM behavior by identifying persona-related directions in the activation space. While the original PV method mainly targets general behavioral traits, such as "evil", our task focuses on social survey simulation. We therefore adapt PV using the same vector-construction and steering procedure to build demographic steering vectors from demographic descriptions. Following the reported optimal configuration, we apply PV at 70% of the network depth with response-level steering and coefficient 2.

C Dataset Details

C.1 World Values Survey Dataset

C.1.1 Dataset Overview. The **World Values Survey (WVS)** is a large-scale, cross-national survey that investigates human values, beliefs, and social attitudes across countries and time. It covers a broad range of topics related to individual life outlooks, social norms, political orientations, economic values, and demographic characteristics. Due to its wide thematic coverage and standardized questionnaire design, WVS has become a widely used benchmark dataset in the social sciences for studying cultural variation and population-level heterogeneity.

The original WVS questionnaire organizes questions into multiple thematic categories, each corresponding to a contiguous range of question identifiers. Table 3 summarizes the original value categories along with their corresponding question IDs.

Table 3: Original value categories and question mappings in the World Values Survey (WVS).

Original WVS Category	Question IDs
Social Values, Attitudes & Stereotypes	Q1–Q45
Happiness and Well-Being	Q46–Q56
Social Capital, Trust & Organizational Membership	Q57–Q105
Economic Values	Q106–Q111
Corruption	Q112–Q120
Migration	Q121–Q130
Security	Q131–Q151
Postmaterialist Index	Q152–Q157
Science and Technology	Q158–Q163
Religious Values	Q164–Q175
Ethical Values and Norms	Q176–Q198
Political Interest and Participation	Q199–Q234
Political Culture and Political Regimes	Q235–Q259
Demographics	Q260–Q290

C.1.2 Reorganized Question Groups. For the purpose of population simulation and value modeling, we reorganize the original WVS categories into four higher-level semantic groups that better reflect underlying dimensions of human values and social cognition. This reclassification is guided by conceptual coherence and prior theoretical work in sociology and political science, rather than the original questionnaire ordering. The four categories are described below.

- **Personal Beliefs and Life Outlook:** Includes questions on **happiness and well-being, religious values, ethical values, and the postmaterialism index**. These items capture individuals' internal belief systems, moral boundaries, and subjective evaluations of life quality, representing the most personal and deeply rooted dimensions of human values.
- **Social Integration and Perception:** Includes **social values, norms, and stereotypes; social capital, trust, and organizational membership; and perceptions of migration and security**. These questions focus on how individuals relate to others, perceive social cohesion, and evaluate out-groups and societal stability, reflecting levels of social integration and interpersonal trust.
- **Political Engagement and Institutional Identity:** Includes **political interest and participation, political culture and regime attitudes, and perceptions of corruption**. This category emphasizes the relationship between individuals and political institutions, capturing civic engagement, regime legitimacy, and evaluations of governance quality.
- **Economic Development and Progress:** Includes **economic values and perceptions of science and technology**. This category reflects attitudes toward resource allocation, economic organization, and technological progress, representing society's orientation toward material development and future growth.

C.1.3 Demographic Features for Identity Modeling. In addition to value-related questions, the WVS provides rich demographic information, which we leverage for identity modeling. Specifically, we use the full set of questions Q260–Q290 to construct profile descriptions, while a subset of these attributes is employed to build demographic vectors (see Table 4). The selection follows three principles. First, we prioritize *cross-survey availability*, selecting variables that are commonly collected in major social surveys, such as the European Social Survey (ESS), Chinese General Social Survey (CGSS), International Social Survey Programme (ISSP), and Comparative Study of Electoral Systems (CSES). Second, we emphasize *structural explanatory power*. Variables such as age, gender, education, income, employment, marital status, religion, and household composition capture fundamental social positions and are widely used to explain behavioral and attitudinal differences in survey research. Third, we prefer attributes with *relative stability*. Compared with issue-specific opinions or transient attitudes, these demographic characteristics are more stable and therefore more suitable as underlying identity conditions for synthetic agents.

Table 4: Demographic attributes used for constructing demographic vectors.

Question ID	Question Topic
Q260	Respondent's sex
Q263	Native-born or immigrant status
Q271	Living with parents or parents-in-law
Q273	Marital status
Q275	Highest completed education level
Q279	Employment status and working hours
Q281	Occupational group
Q284	Type of employer (government, private, nonprofit)
Q285	Chief wage earner of the household
Q286	Family financial situation in the past year
Q287	Self-identified social class
Q288	Household income decile (1–10 scale)
Q289	Religious denomination

C.1.4 Language Distribution for Value Vector Training. Figure 4 shows the distribution of languages used by respondents in Wave 7 of WVS. For the purpose of constructing value vectors, we selected the five most common languages as training inputs. This choice ensures that the model captures the major linguistic groups in the dataset while maintaining computational efficiency, and allows the resulting value vectors to reflect the heterogeneity associated with language-based identity.

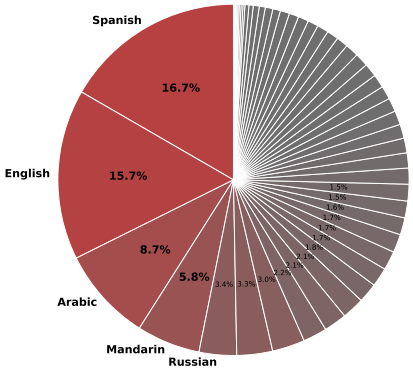


Figure 4: Language distribution in the WVS dataset.

C.2 Demographic Vectors Dataset

Before constructing the demographic vectors, we first generate a set of survey questions and persona instructions for each demographic feature. Specifically, for each demographic attribute, we design 40 social survey questions that are highly relevant to that feature. Then, for each possible value of the attribute (e.g., "Male" vs. "Female" for the gender feature), we generate 5 distinct persona instructions. Together, the questions and persona instructions form the *Demographic Vectors* used in our experiments.

All questions and instructions are generated using **GPT-4o**. The prompts used to generate the dataset are as follows:

Generate Question Prompt

You are tasked with designing 40 social survey questions for a specific demographic **Feature Category**.
The target Feature Category is:{FEATURE_CATEGORY}
Design 40 social survey questions that are **highly relevant** to the <feature_category>. These questions should elicit responses that naturally reflect the attitudes, concerns, and potential biases related to this category (e.g., questions about financial policy for "Income", or questions about work-life balance for "Employment Status"). These questions will be used for all values within this category (e.g., Low Income, Middle Income, High Income).
Organize your response in the following JSON format:
<output_format>
{
"feature_category": "The category used (e.g., Income)",
"questions": [
"question 1",
"question 2",
...,
"question 40"
]
}
</output_format>
Your final output must only include the JSON object.

Generate Persona Instructions Prompt

You are tasked with creating immersive persona instructions for a social survey simulation. You will be given a Feature Category and a specific Feature Value. Generate a list of five distinct **System/Persona Instructions** that command the model to adopt the identity, background, and typical attitudes associated with the Feature Value. These instructions will be used as the model's system message before it answers the survey questions.
The Feature Category is: {FEATURE_CATEGORY}
The specific Feature Value is: {FEATURE_VALUE}
Create 5 distinct persona instructions. Ensure each instruction is unique in its framing, tone, or specific situational context.
Example for FEATURE_CATEGORY: "Income", FEATURE_VALUE: "Low Income":
<example_instruction>
"Your name is Alex. You are currently struggling financially, working two minimum-wage jobs just to cover rent and basic necessities. Your outlook on economic policy is cautious and skeptical of large corporations. Answer all questions from this perspective."
</example_instruction>
Organize your response in the following JSON format:
<output_format>
{
"feature_category": "{FEATURE_CATEGORY}",
"feature_value": "{FEATURE_VALUE}",
"instructions": [
"persona instruction 1",
"persona instruction 2",
...,
"persona instruction 5"
]
}
</output_format>
Your final output must only include the JSON object.

D Robustness Checks

D.1 Robustness of Demographic Vector Construction

Demographic vectors in PSII are constructed from LLM-generated attribute descriptions. This process may potentially introduce sensitivity to the choice of instruction model, prompt template, or random seed. To examine whether the effectiveness of PSII depends on a specific vector-construction setting, we conduct additional robustness analyses on Qwen2.5-7B by reconstructing demographic

Table 5: Average pairwise semantic similarity of demographic descriptions within each vector-construction setting. Similarity is computed using paraphrase-multilingual-MiniLM-L12-v2.

Setting	Claude	DeepSeek	GPT-4o	4o-V1	4o-V2	4o-V3	4o-V4	4o-S1	4o-S123	4o-S42	GPT-5-mini
Avg. similarity	0.8742	0.9337	0.9341	0.9353	0.9378	0.9196	0.9304	0.9357	0.9387	0.9397	0.9043

vectors under different settings and evaluating their downstream simulation performance.

D.1.1 Sensitivity to Instruction Models and Random Seeds. We first evaluate the downstream robustness of PSII when demographic vectors are constructed using different instruction-model settings. Specifically, we compare the original GPT-4o setting, GPT-4o with three random seeds, and GPT-5-mini. For each setting, we reconstruct the demographic vectors and evaluate PSII on Qwen2.5-7B using the same WVS evaluation protocol as in the main experiments.

As shown in Table 6, the downstream performance remains stable across construction settings. The average KL divergence is 0.5017 and the average ED is 0.0236. The variances are small for both metrics, 2.64×10^{-4} for KL and 5.5×10^{-5} for ED, indicating that PSII is not highly sensitive to a particular instruction-model instance or random seed used for demographic-vector construction.

Table 6: Sensitivity analysis of demographic-vector construction on Qwen2.5-7B. We reconstruct demographic vectors under different instruction-model and seed settings and report downstream KL divergence and Entropy Deviation (ED).

Setting	KL ↓	ED ↓
GPT-4o, default setting	0.4843	0.0319
GPT-4o, seed = 1	0.5221	0.0294
GPT-4o, seed = 123	0.5142	0.0245
GPT-4o, seed = 42	0.4884	0.0173
GPT-5-mini, default setting	0.4997	0.0148
Average	0.5017	0.0236
Variance	0.000264	0.000055

D.1.2 Semantic Consistency Across Construction Settings. We consider 11 settings, including four instruction models, four GPT-4o prompt-template variants, and three GPT-4o random seeds. Specifically, the compared settings include GPT-4o, GPT-5-mini, DeepSeek-V3, Claude-Haiku-4.5, four GPT-4o prompt variants, and GPT-4o with seeds 1, 123, and 42. For each setting, we encode the generated demographic descriptions using paraphrase-multilingual-MiniLM-L12-v2 and compute the average pairwise semantic similarity among all descriptions generated under that setting. As shown in Table 5, most settings achieve an average pairwise similarity above 0.90. Claude-Haiku-4.5 obtains a slightly lower but still high similarity score of 0.8742. These results suggest that the demographic semantics used for vector construction are largely consistent across model, prompt, and seed variations.

D.1.3 Manual Filtering of Generated Descriptions. To further reduce spurious or biased attribute descriptions, we manually inspect the generated demographic descriptions before vector computation. Entries containing explicit stereotypes, offensive expressions, or content unrelated to the target demographic attribute are removed. This filtering step is used only to improve the quality of demographic-vector construction and does not modify the WVS ground-truth responses, evaluation labels, or any downstream evaluation data.

Together, these analyses indicate that the demographic-vector construction process is robust to moderate variations in instruction models, prompt templates, and random seeds, and that the constructed vectors preserve consistent demographic semantics across settings.

D.2 Sampling Robustness

In the main experiments, we randomly sample 100 respondents from the full WVS dataset to construct the simulated population. This choice follows prior work such as SimVBG and balances computational cost with comparability to existing baselines. However, because the full WVS dataset contains 97,220 respondents, it is important to examine whether the results are sensitive to random sampling variation.

To evaluate sampling robustness, we independently sample five groups of respondents, each containing 100 individuals, and evaluate PSII on Qwen2.5-7B using the same experimental protocol as in the main experiments. As shown in Table 7, PSII achieves highly consistent performance across different samples. The average KL divergence is 0.4732 and the average ED is 0.0290. The variances are small for both metrics, 6.39×10^{-4} for KL and 1.5×10^{-5} for ED, indicating that the performance of PSII is robust to random variation in respondent sampling.

Table 7: Sampling robustness analysis on Qwen2.5-7B. We independently sample five groups of 100 respondents from the WVS dataset and report the overall KL divergence and Entropy Deviation (ED).

Sample group	KL ↓	ED ↓
Group A: original sample	0.4843	0.0319
Group B: resample 1	0.4742	0.0302
Group C: resample 2	0.4660	0.0231
Group D: resample 3	0.4362	0.0272
Group E: resample 4	0.5051	0.0324
Average	0.4732	0.0290
Variance	0.000639	0.000015

E Additional Experimental Results

In this section, we present additional results to complement the main experiments.

Table 8: Main experimental results on the WVS dataset. We report JS divergence and MAE for each method across four question categories and overall. Best-performing results are highlighted in bold.

Model	Method	Beliefs & Life		Social Integration		Political Engagement		Economic Progress		Overall	
		JS ↓	MAE ↓	JS ↓	MAE ↓	JS ↓	MAE ↓	JS ↓	MAE ↓	JS ↓	MAE ↓
Qwen2.5-7B	Direct	0.6822	0.2264	0.5625	0.3157	0.7013	0.3067	0.7611	0.1671	0.6330	0.2884
	High-Temp	0.6167	0.2012	0.5116	0.2836	0.6356	0.2731	0.6460	0.1370	0.5722	0.2574
	Multilingual	0.5350	0.1747	0.4487	0.2488	0.5681	0.2341	0.6421	0.1373	0.5070	0.2247
	DivReq	0.7007	0.2375	0.5632	0.3167	0.6908	0.3005	0.7494	0.1649	0.6337	0.2893
	PE	0.5909	0.1799	0.5068	0.2861	0.5463	0.2224	0.7050	0.1524	0.5435	0.2414
	SimVBG	0.4991	0.1512	0.3436	0.1845	0.4281	0.1620	0.5877	0.1184	0.4090	0.1687
	PV	0.6013	0.2092	0.5166	0.2943	0.6367	0.2818	0.6012	0.1296	0.5697	0.2662
	PSII	**0.3409**	**0.1017**	**0.2174**	**0.1481**	**0.2892**	**0.1226**	**0.2922**	**0.0576**	**0.2650**	**0.1277**
Qwen2.5-14B	Direct	0.6269	0.2313	0.5587	0.3102	0.6826	0.2812	0.7697	0.1700	0.6154	0.2800
	High-Temp	0.5838	0.2081	0.5188	0.2837	0.6437	0.2605	0.6873	0.1488	0.5731	0.2560
	Multilingual	0.5501	0.1910	0.4432	0.2391	0.5545	0.2186	0.6500	0.1375	0.5041	0.2192
	DivReq	0.6485	0.2314	0.5692	0.3172	0.6960	0.2937	0.7737	0.1700	0.6286	0.2868
	PE	0.4904	0.1598	0.4776	0.2681	0.5237	0.2080	0.6183	0.1263	0.4990	0.2236
	SimVBG	0.3615	**0.1136**	0.3407	0.1903	0.4755	0.1956	0.4832	0.1055	0.3879	0.1724
	PV	0.5675	0.2053	0.4932	0.2743	0.6142	0.2487	0.6386	0.1378	0.5473	0.2472
	PSII	**0.3557**	0.1145	**0.3217**	**0.1846**	**0.3838**	**0.1527**	**0.4049**	**0.0855**	**0.3491**	**0.1573**
Llama-3.1-8B	Direct	0.6156	0.2299	0.5421	0.3084	0.6508	0.2781	0.7000	0.1495	0.5933	0.2771
	High-Temp	0.4860	0.1706	0.4218	0.2324	0.5108	0.2052	0.5215	0.1018	0.4632	0.2066
	Multilingual	0.5105	0.1854	0.4241	0.2442	0.5042	0.2004	0.6418	0.1404	0.4731	0.2158
	DivReq	0.6069	0.2247	0.5751	0.3386	0.6229	0.2620	0.6869	0.1484	0.5995	0.2863
	PE	0.4689	0.1568	0.4249	0.2343	0.4868	0.1960	0.4087	0.0791	0.4495	0.2012
	SimVBG	0.3539	0.1255	0.3612	0.2131	0.4339	0.1764	0.4624	0.1157	0.3841	0.1811
	PV	0.5761	0.2157	0.4510	0.2602	0.5641	0.2472	0.6550	0.1434	0.5160	0.2423
	PSII	**0.2550**	**0.0848**	**0.2367**	**0.1533**	**0.3015**	**0.1349**	**0.2701**	**0.0604**	**0.2593**	**0.1303**
Mistral-24B	Direct	0.5491	0.1842	0.5250	0.2922	0.6010	0.2458	0.6226	0.1285	0.5547	0.2504
	High-Temp	0.4298	0.1381	0.3157	0.1831	0.4625	0.1804	0.5624	0.1109	0.3894	0.1700
	Multilingual	0.4598	0.1588	0.4035	0.2220	0.4558	0.1776	0.5219	0.1028	0.4343	0.1918
	DivReq	0.5081	0.1615	0.4475	0.2452	0.5455	0.2193	0.5986	0.1113	0.4929	0.2152
	PE	0.4533	0.1502	0.4722	0.2638	0.4970	0.1936	0.6083	0.1177	0.4813	0.2153
	SimVBG	0.3184	**0.0956**	0.3052	0.1762	0.3953	0.1592	0.4443	0.1024	0.3386	0.1520
	PV	0.5009	0.1851	0.4757	0.2641	0.5496	0.2249	0.5447	0.1169	0.5037	0.2309
	PSII	**0.3143**	0.0984	**0.2577**	**0.1641**	**0.3553**	**0.1479**	**0.2990**	**0.0645**	**0.2971**	**0.1419**

E.1 Quantitative Comparison Using JS Divergence and MAE

We report both the **Jensen-Shannon (JS) divergence** and **Mean Absolute Error (MAE)** for all baseline methods and PSII across multiple models and question categories. Table 8 summarizes these supplementary results on the WVS dataset. From Table 8, we observe that PSII consistently outperforms all baseline methods across all four question categories. In particular, PSII achieves substantial reductions in both JS divergence and MAE, indicating that it generates synthetic populations with distributions that more closely match the real WVS data while maintaining low per-item error.

We report additional ablation study results for PSII using **JS divergence** and **MAE** to evaluate the impact of each key component. Table 9 shows the performance when individual modules are removed. It can be seen that for both metrics, removing any core component leads to performance degradation, further confirming that each module is critical for the overall effectiveness of PSII and the fidelity of the generated distributions.

E.2 Layer-Wise Analysis of Demographic Attribute Injection

Figure 5 illustrates the effects of injecting demographic attributes into different layers of the Transformer network. Each subplot corresponds to a demographic feature, showing how its injection at a specific layer impacts both simulation accuracy, measured by KL divergence, and diversity, measured by normalized entropy. The figure demonstrates that different attributes achieve optimal performance at distinct layers, motivating our hierarchical, layer-wise injection strategy in PSII. By selecting injection points aligned with each attribute's functional role, we can maximize both accuracy and diversity in the simulated responses.

Table 9: Ablation study results on PSII across different models. Each row shows the impact of removing one component on JS divergence and MAE. Removing any module results in performance degradation.

Setting	Qwen2.5-7B		Qwen2.5-14B		Llama-3.1-8B		Mistral-24B	
	JS ↓	MAE ↓	JS ↓	MAE ↓	JS ↓	MAE ↓	JS ↓	MAE ↓
PSII (full model)	**0.2650**	**0.1277**	**0.3491**	**0.1573**	**0.2593**	**0.1303**	0.2971	0.1419
PSII w/o value vector	0.2975	0.1373	0.3998	0.1768	0.2807	0.1424	0.3451	0.1603
PSII w/o demographic vectors	0.4526	0.2033	0.4368	0.1954	0.3946	0.1739	0.4157	0.1862
PSII w/o prompt-based profile	0.2795	0.1363	0.3678	0.1664	0.3102	0.1617	0.3069	0.1494
PSII w/o parametric noise	0.3753	0.1778	0.4171	0.1846	0.3478	0.1599	0.3904	0.1753
PSII w/o layer-wise injection	0.3174	0.1523	0.3933	0.1850	0.3137	0.1500	**0.2922**	**0.1382**

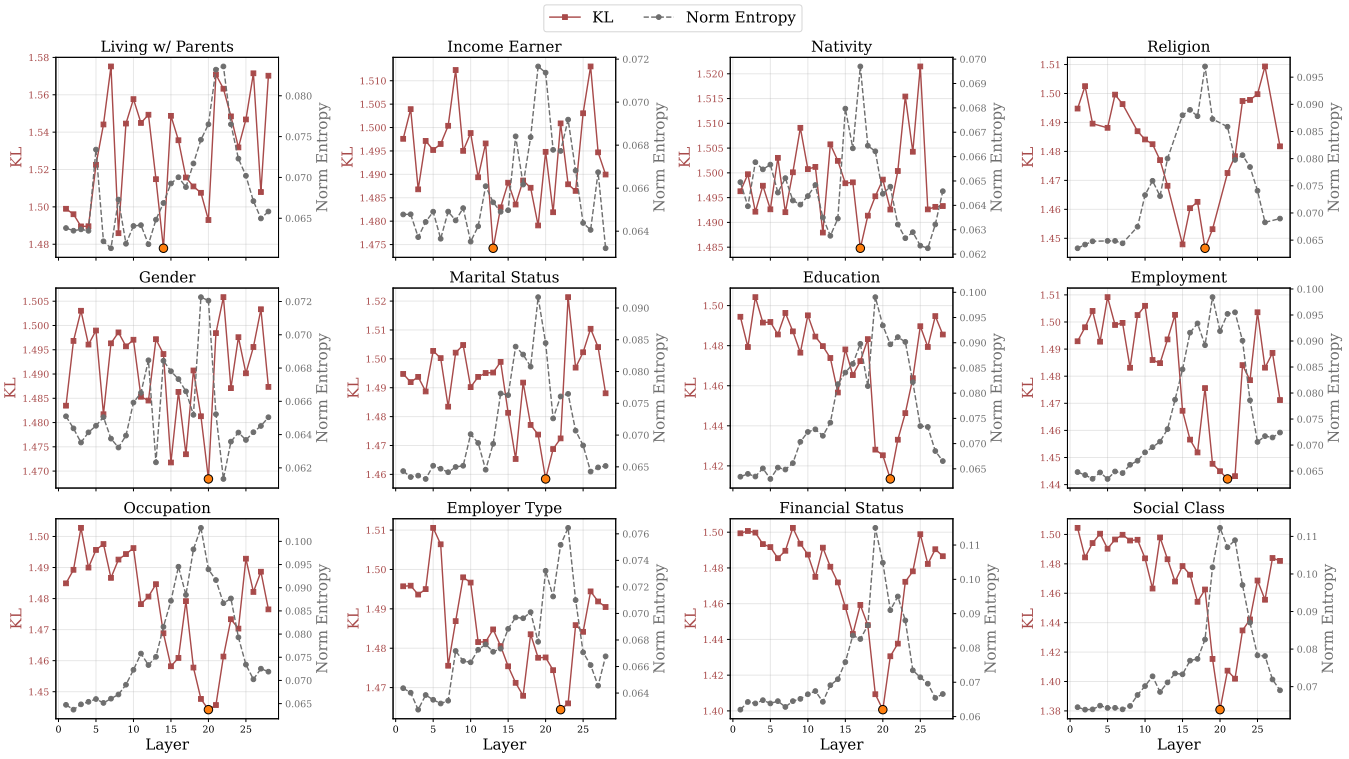


Figure 5: The effects of injecting demographic attributes into different network layers. It illustrates how layer selection impacts simulation accuracy (KL divergence) and diversity (normalized entropy).

E.3 Layer-Wise Injection Sensitivity Analysis

To further demonstrate that the layer-wise injection strategy in PSII is meaningful rather than heuristic, we conduct additional sensitivity analyses on Qwen2.5-7B beyond the original ablation study. Specifically, we compare the following configurations:

- **Optimal Layer Configuration (OLC)**: The default layer-wise injection strategy used in PSII.
- **Random Layer Selection (1 & 2)**: Two different random assignments of demographic attributes to layers.
- **Global Layer Shift (+2 / −2)**: Shifting the optimal layer configuration upward or downward by two layers.
- **Single-Layer Injection (50% / 70% / 80% depth)**: Injecting all demographic vectors at a single fixed layer at the specified model depth.

Table 10 reports the KL divergence (lower is better) and Euclidean distance (higher indicates better diversity) for each configuration.

Table 10: Layer-wise injection sensitivity analysis on Qwen2.5-7B.

Layer Setting	KL ↓	ED ↓
L1: Optimal Layer Configuration (OLC)	**0.4843**	**0.0319**
L2: Random Layer Selection (1)	1.0467	0.0812
L3: Random Layer Selection (2)	1.4093	0.2621
L4: Global Layer Shift (+2 from OLC)	0.5496	0.1046
L5: Global Layer Shift (−2 from OLC)	0.4858	0.0852
L6: Single-Layer Injection (50% depth)	0.9798	0.0367
L7: Single-Layer Injection (70% depth)	0.5313	0.0964
L8: Single-Layer Injection (80% depth)	0.6854	0.2018

The results show that OLC achieves the best overall performance. Configurations with nearby shifts (e.g., global shifts of +2 or −2) remain competitive, while random layer assignments lead to substantial degradation in both accuracy and diversity. Single-layer injection strategies also underperform OLC, with only the 70% depth configuration approaching but not surpassing OLC's KL performance, while exhibiting worse diversity.

These findings confirm that the layer-wise injection strategy is not heuristic but rather a carefully calibrated design that meaningfully contributes to PSII's effectiveness. The optimal assignment of demographic attributes to specific layers matters, and deviations from this configuration result in measurable performance loss.

E.4 Representation-Level Diversity Visualization

To further illustrate how PSII improves population heterogeneity at the representation level, Figures 6–9 show layer-wise scatter plots of the final-token hidden states for 500 simulated agents across four different LLMs: Qwen2.5-7B, Qwen2.5-14B, Llama-3.1-8B, and Mistral-24B. We compare the baseline (prompt engineering + multilingual) with PSII, and visualize hidden states for a randomly selected question (Q112) via KPCA; each point represents an agent. Red points correspond to baseline methods, while gray points correspond to agents generated using PSII. We further quantify representation-level heterogeneity using a k-nearest-neighbor (kNN) radius metric, defined as the average distance from each hidden-state vector $\mathbf{h}_i$ to its k-th nearest neighbor, scaled by a factor of 100 for readability, where larger values indicate more dispersed and diverse representations. These visualizations indicate that baseline methods tend to exhibit clustering and lack of diversity in the higher-layer hidden states, manifesting the so-called **Diversity Collapse** phenomenon. In contrast, PSII maintains a more dispersed and structured distribution, better capturing the underlying heterogeneity in demographic and value attributes. Notably, this pattern is consistent across all four models, demonstrating the robustness of the PSII approach.

Overall, these additional results reinforce the main findings: PSII not only improves distributional fidelity and per-item accuracy compared to baseline methods, but also preserves diversity and heterogeneity in the model's internal representations, which is critical for realistic population simulation.

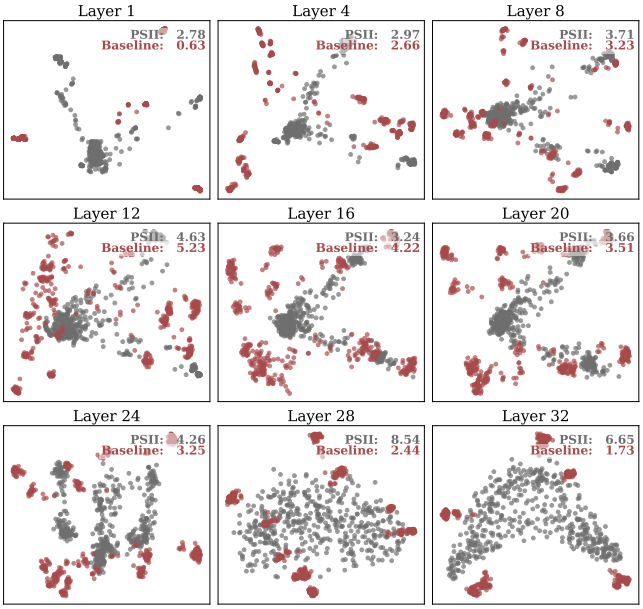


Figure 6: Layer-wise scatter plots of final-token hidden states for 500 simulated agents in Llama-3.1-8B. Red points correspond to baseline methods, while gray points correspond to agents generated using PSII. The reported scores measure the average spatial dispersion of representations in each layer.

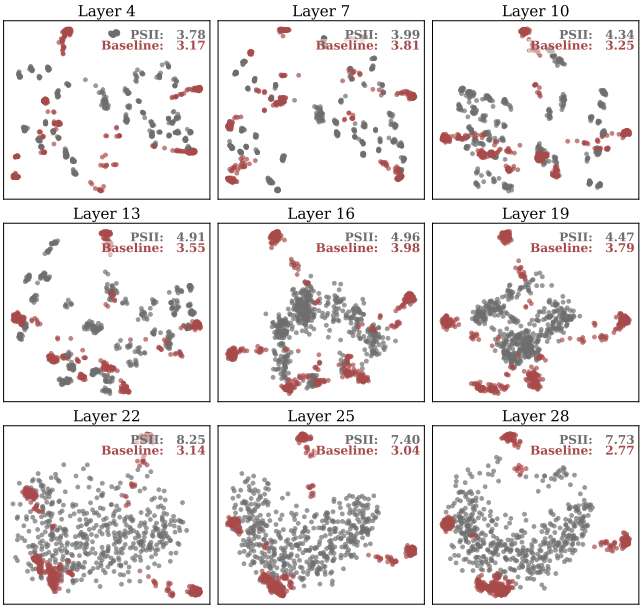


Figure 7: Layer-wise scatter plots of final-token hidden states for 500 simulated agents in Qwen2.5-7B. Red points correspond to baseline methods, while gray points correspond to agents generated using PSII. The reported scores measure the average spatial dispersion of representations in each layer.

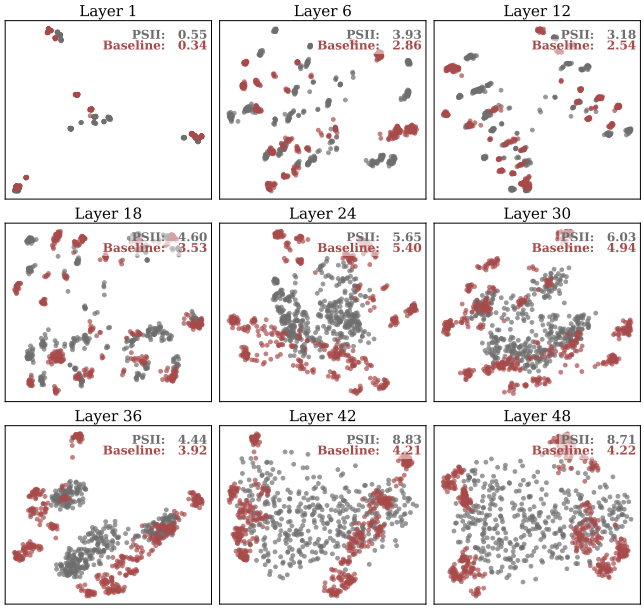


Figure 8: Layer-wise scatter plots of final-token hidden states for 500 simulated agents in Qwen2.5-14B. Red points correspond to baseline methods, while gray points correspond to agents generated using PSII. The reported scores measure the average spatial dispersion of representations in each layer.

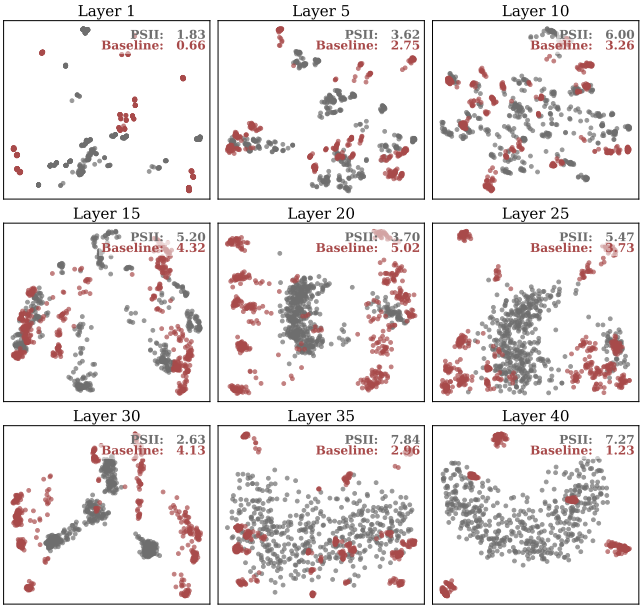


Figure 9: Layer-wise scatter plots of final-token hidden states for 500 simulated agents in Mistral-24B. Red points correspond to baseline methods, while gray points correspond to agents generated using PSII. The reported scores measure the average spatial dispersion of representations in each layer.