HAG: Hierarchical Demographic Tree-based Agent Generation for Topic-Adaptive Simulation

Rongxin Chen[1,2,3], **Tianyu Wu**[1,2,3], **Bingbing Xu**[1,2*],
Jiatang Luo[1,2,4], **Xiucheng Xu**[1,2,3], **Huawei Shen**[1,2,3]
[1]State Key Laboratory of AI Safety, Beijing, 100086
[2]Institute of Computing Technology, Chinese Academy of Sciences
[3]University of Chinese Academy of Sciences
[4]School of Advanced Interdisciplinary Sciences, University of Chinese Academy of Sciences
{chenrongxin24s,wutianyu25s,xubingbing}@ict.ac.cn

Abstract

High-fidelity agent initialization is crucial for credible Agent-Based Modeling across diverse domains. A robust framework should be Topic-Adaptive, capturing macro-level joint distributions while ensuring micro-level individual rationality. Existing approaches fall into two categories: static data-based retrieval methods that fail to adapt to unseen topics absent from the data, and LLM-based generation methods that lack macro-level distribution awareness, resulting in inconsistencies between micro-level persona attributes and reality. To address these problems, we propose HAG, a Hierarchical Agent Generation framework that formalizes population generation as a two-stage decision process. Firstly, utilizing a World Knowledge Model to infer hierarchical conditional probabilities to construct the Topic-Adaptive Tree, achieving macro-level distribution alignment. Then, grounded real-world data, instantiation and agentic augmentation are carried out to ensure micro-level consistency. Given the lack of specialized evaluation, we establish a multi-domain benchmark and a comprehensive PACE evaluation framework. Extensive experiments show that HAG significantly outperforms representative baselines, reducing population alignment errors by an average of 37.7% and enhancing sociological consistency by 18.8%. Code and dataset are released at https://github.com/Libra117/HAG.

1 Introduction

With the advancement of Large Language Models (LLMs), agent-based modeling (ABM) plays a key role in simulating complex multi-agent interactions, with growing adoption in computational social science (Gao et al., 2024; Manning et al., 2024; Wang et al.), economic modeling (Li et al., 2024b; Horton, 2023), and personalized recommendation (Zhu et al., 2025; Peng et al., 2025). These simulations share a common requirement: heavily depend on User Agents to simulate preferences and interactions. Therefore, the fidelity of such a simulation system is essentially limited by the quality of agents, whether the simulated population accurately represents the demographic heterogeneity and behavioral differences of real world (Wang et al., 2025d; Bui et al., 2025).

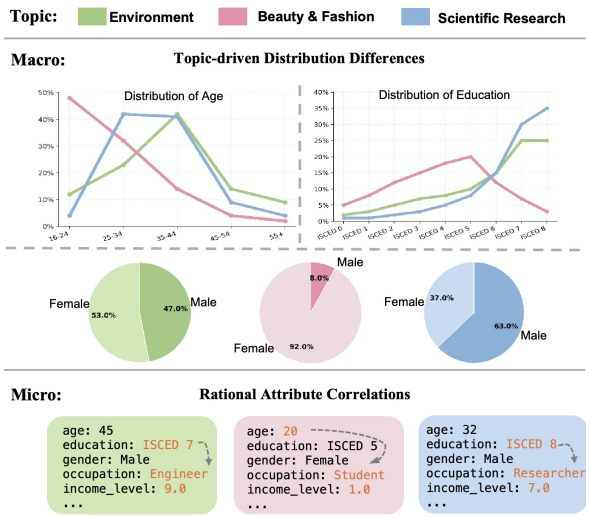

Figure 1: The necessity of Topic-Adaptive in simulation.

A robust agent generation framework for simulation should be fundamentally Topic-Adaptive. As illustrated in Figure 1, demographic structures and attribute distributions are not static but changing across different topics. Capturing these topic-specific differences requires satisfying two critical criteria: at the macro level, the generated population should model the correct joint distribution of multi-dimensional attributes conditioned on the topic, ensuring the overall demographic persona aligns with the specific scenario rather than treating agents as isolated individuals; and at the micro level, individual agents should present rational and coherent attribute correlations, reflecting realistic sociological dependencies rather than random com-

* Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 25890–25906

binations. Failure to meet these criteria inevitably leads to significant divergence in simulation trajectories and unreliable downstream analysis.

Existing agent generation approaches (Wang et al., 2024; Chen et al., 2024; Tseng et al., 2024), however, struggle to fulfill these criteria. They fall into two categories: data-based retrieval and LLM-based generation. Data-based retrieval constructs agent pools directly from real-world user logs (Zhang et al., 2025; Yang et al., 2024; Wang et al., 2025a; Jansen et al., 2022), but are inherently static and tightly coupled to historical data, limiting their ability to adapt to unseen or data-scarce topics. LLM-based generation methods construct agent personas via predefined schemas (Wang et al., 2025b; Li et al., 2025; Schuller et al., 2024) or text-based inference (Ge et al., 2024; Hu et al., 2025). While more flexible, they are typically constrained by sparse expert knowledge and construct population by aggregating independent individuals, lacking explicit modeling of the joint distributions over multi-dimensional attributes and multi-agents (Xie et al., 2025; Madden, 2025). As a result, when lacking real data support, individuals may exhibit incongruous persona attributes, rendering simulations ineffective (Pal and Traum, 2025; Larooij and Törnberg, 2025; Li et al., 2025; Feng et al., 2025). Overall, to date, no existing approach simultaneously achieves both topic-adaptive population macro-level modeling and micro-level rationality.

To combat the challenges, we propose **HAG**, a **H**ierarchical **A**gent **G**eneration framework that formalizes population generation as a hierarchical decision process. By encoding the target topic as a latent guiding prompt, HAG achieves progressive persona generation in two stages: 1) Topic-Adaptive Tree Construction: Utilizing a World Knowledge Model, the framework autonomously derives node values and path weights by computing hierarchical conditional probabilities across salient demographic dimensions. This replaces manual heuristics with automated expert priors, ensuring macro-distributions are contextually aligned with the target topic. 2) Grounded Instantiation & Agentic Augmentation: To populate the tree, the framework grounds leaf nodes via real-world persona retrieval. For data-deficient nodes, we employ agentic augmentation to synthesize missing personas, satisfying global macro-constraints while maintaining micro-level consistency. In summary, HAG enables the efficient generation of diverse, context-aware, and sociologically grounded agent populations for high-fidelity simulations.

Given the lack of specialized evaluation for agent generation, we establish a multi-domain benchmark, spanning social simulation, product recommendation, and movie critique, and we propose **PACE** (**P**opulation **A**lignment & **C**onsistency **E**valuation). This framework provides comprehensive metrics to quantitatively assess the generated population from two complementary dimensions: statistical alignment and sociological consistency. Extensive experiments show that HAG significantly outperforms representative baselines, reducing population alignment errors by an average of **37.7%** and enhancing sociological consistency by **18.8%**. Code and benchmark are released.

2 Methodology

In this section, we propose **HAG**, a **H**ierarchical Demographic Tree-based **A**gent **G**eneration framework for Topic-Adaptive Simulation. We first provide a formal definition of the topic-adaptive population generation problem, and then detail the HAG framework. An illustration of our framework is shown in the left part of Figure 2.

2.1 Problem Definition

We consider the task of constructing a topic-based agentic population dataset grounded in real demographic data. For a given input topic t and a target population size N, our goal is to generate an agent population $\mathcal{P} = \{p_j\}_{j=1}^{N}$ whose demographic composition matches the target demographic structure of that topic. Each agent p_j is associated with a persona vector $\mathbf{x}_j = \{v^{(1)}, v^{(2)}, \ldots, v^{(L)}\}$ defined over a set of L selected demographic dimensions (e.g., Age, Gender), and each $v^{(i)}$ represents the specific value of the agent on the i-th dimension.

The core challenge lies in the fact that the ideal demographic composition for a specific topic is often latent. We denote this target persona distribution as $\mathbf{D}(t)$, which represents the distribution of demographic attributes given the topic t. At the macro level, the real distribution of the generation dataset, denoted as $\widehat{\mathbf{D}}_{\mathcal{P}}$, should approximate the theoretical target $\mathbf{D}(t)$, capturing the correct conditional dependencies between attributes rather than treating them independently. At the micro level, each generated individual $\mathbf{x}_j$ should fall within the valid real-world data, avoiding unrealistic attribute combinations. Consequently, our objective is to derive a structured representation of $\mathbf{D}(t)$ and in-

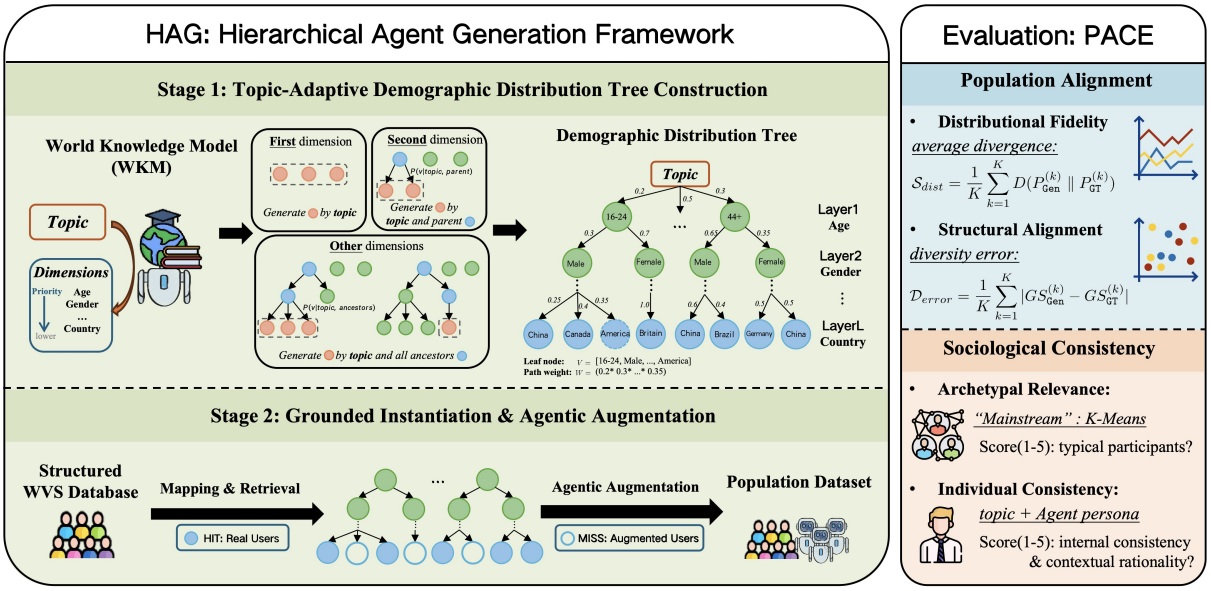

Figure 2: Illustration of the HAG framework. It utilizes the World Knowledge Model to construct a Topic-Adaptive Demographic Distribution Tree, and generates population based on the tree via filtering real users and agentic data augmentation. Evaluation of the generated population is from two aspects: population alignment and sociological consistency.

stantiate a concrete population dataset that satisfies both topic-adaptivity and grounding constraints.

2.2 HAG Framework

The HAG framework achieves this objective through two stages: (1) constructing a hierarchical demographic distribution tree to capture topic-relevant demographic dimensions and their joint probabilities of values. (2) grounding this tree in real-world data by filtering real users and augmenting generated agents where data is missing. This effectively bridges the gap between macro-level distribution modeling and micro-level individual consistency.

2.2.1 Topic-Adaptive Demographic Distribution Tree Construction

To explicitly model the macro-level joint distribution, we represent the topic-based population structure as a hierarchical demographic distribution tree $\mathcal{T}_d$.

For a given topic t, we use a World Knowledge Model (WKM) to identify and order a subset of relevant demographic dimensions from a pre-specified set $\mathcal{F}$. Formally, the WKM outputs a prioritized dimension sequence:

$$\begin{aligned}\mathbf{f}_{\text{prior}}(t) &= \text{PrioritizeDims}(\mathcal{M},\, t,\, \mathcal{F}) \\ &= \left(f^{(1)}, \ldots, f^{(L)}\right)\end{aligned} \tag{1}$$

where $\mathcal{M}$ denotes the WKM, $\text{PrioritizeDims}(\cdot)$ is the function to identify and order topic-relevant dimensions. This sequence directly determines the layer order of the distribution tree from root to leaves, where $f^{(1)}$ represents the first dimension with the highest priority and $f^{(L)}$ has the lowest priority among the selected dimensions.

The tree is constructed top-down along this ordered sequence driven by the WKM. With the input topic t as the root node, the demographic dimensions are encoded in hierarchical order of priority, with higher priority dimensions expanding earlier. At the first layer, nodes are generated directly conditioned on the topic t. For any subsequent layer l, nodes are generated based on both the topic and the path of ancestor nodes. Specifically, each node represents a distinct value $v^{(l)}$ of dimension $f^{(l)}$, and each edge encodes a conditional weight. For a parent node representing a partial persona $\left(v^{(1)}, \ldots, v^{(l-1)}\right)$, the normalized weight of the edge to a child value $v^{(l)}$ is defined as:

$$\begin{aligned}&w\left(v^{(l)} \mid v^{(1:l-1)}, t\right) \\ =&P\left(f^{(l)} = v^{(l)} \mid f^{(1:l-1)} = v^{(1:l-1)}, t\right)\end{aligned} \tag{2}$$

where $\sum_{v^{(l)} \in \mathcal{V}(f^{(l)})} w\left(v^{(l)} \mid v^{(1:l-1)}, t\right) = 1$. This formulation ensures that the probability of each attribute is not static but dynamically conditioned on the preceding context, capturing the dependencies between dimensions.

Category	Dimension	WVS Code
Basic Demographics	Country	B_COUNTRY
	Language	S_INTLANGUAGE
	Gender	Q260
	Age	Q262
	Marital Status	Q273
Socio-Economic Status	Education	Q275
	Occupation	Q281
	Income Level	Q288
	Financial Status	Q286
	Social Class	Q287
Cultural Identity	Religion	Q289
	Ethnicity	Q290

Table 1: The schema of demographic attributes extracted from the WVS dataset.

Finally, each leaf node corresponds to a complete demographic persona $\mathbf{v} = (v^{(1)}, \ldots, v^{(L)})$. The target proportion for this persona is obtained by the product of edge weights along the root-to-leaf path:

$$W(\mathbf{v} \mid t) = \prod_{l=1}^{L} w\big(v^{(l)} \mid v^{(1:l-1)}, t\big) \qquad (3)$$

Each $W(\mathbf{v} \mid t)$ shows the proportion of the population with persona $\mathbf{v}$ under topic t. The resulting tree $\mathcal{T}_d$ provides an explicit, structured macro-constraint for population instantiation, translating the abstract topic into a concrete target joint distribution.

2.2.2 Grounded Instantiation and Agentic Augmentation

To ensure the generated agents match the distribution of the real-world society rather than hallucinated statistics, we utilize the **World Values Survey (WVS)**[1] as a source of our database. Following the demographic framework established in *WorldValuesBench* (Zhao et al., 2024), we extract a comprehensive set of attributes to measure a user's sociological positioning. Concretely, we filter the raw survey data to select 12 key dimensions classified into three categories: Basic Demographics, Socio-Economic Status, and Cultural Identity. Shown in Table 1, this classification captures the social attributes of a persona, enabling us to model complex social distributions rooted in real users.

Based on the demographic distribution tree, we construct the final dataset $\mathcal{P}$ by grounding each leaf distribution in this real-world data. Given a total target size N, we compute the required count for each leaf persona as $n(\mathbf{v}) = \mathrm{Round}\big(N \cdot W(\mathbf{v} \mid t)\big)$, where $\mathrm{Round}(\cdot)$ is the rounding function, with minor adjustments to preserve the sum N. We then perform a retrieval-and-augmentation process:

For each leaf persona $\mathbf{v}$, we filter the processed WVS database to retrieve matching real users. Let $m(\mathbf{v})$ denote the count of available real users. We assign a coverage status $\mathrm{Tag}(\mathbf{v})$:

$$\mathrm{Tag}(\mathbf{v}) = \begin{cases} \texttt{HIT}, & m(\mathbf{v}) \geq n(\mathbf{v}), \\ \texttt{MISS}, & m(\mathbf{v}) < n(\mathbf{v}). \end{cases} \qquad (4)$$

For HIT nodes, we sample $n(\mathbf{v})$ real users directly, ensuring maximum realism. For MISS nodes, where data scarcity occurs, we sample all available $m(\mathbf{v})$ and generate the remaining deficiency $n(\mathbf{v}) - m(\mathbf{v})$ via agentic augmentation. Crucially, this augmentation is constrained by the specific path $\mathbf{v}$ defined by the tree, preventing the LLM from hallucinating incompatible attribute combinations.

Finally, we obtain the final agentic population dataset $\mathcal{P}$ consisting of filtered real-world users and augmented agents, matching the distribution tree. The detailed procedure is presented in Algorithm 1. In summary, HAG effectively solves the mode collapse and "Frankenstein" agent problems

Algorithm 1 HAG algorithm

Require: Topic t, Wold Knowledge Model $\mathcal{M}$, Database $\mathcal{D}$, Size N, Dimensions $\mathcal{F}$
Ensure: Agent Population $\mathcal{P}$

Stage 1: Topic-Adaptive Tree Construction

$\mathbf{f} \leftarrow \textsc{PrioritizeDims}(\mathcal{M}, t, \mathcal{F})$
Initialize tree $\mathcal{T}$ with root r
function $\textsc{Expand}(u, l, \mathbf{v}_{path})$
  if $l > |\mathbf{f}|$ **then return**
  $\{(v, w)\} \leftarrow \textsc{Infer}(\mathcal{M}, t, \mathbf{f}[l] \mid \mathbf{v}_{path})$
  for $(v_i, w_i) \in \{(v, w)\}$ **do**
    $c \leftarrow \textsc{Node}(v_i)$
    $\textsc{AddEdge}(u \rightarrow c, w_i)$
    $\textsc{Expand}(c, l{+}1, \mathbf{v}_{path} \cup \{v_i\})$
$\textsc{Expand}(r, 1, \emptyset)$

Stage 2: Grounded Instantiation

$\mathcal{P} \leftarrow \emptyset$
for each leaf persona $\mathbf{v}$ in $\mathcal{T}$ **do**
  $n_{tgt} \leftarrow \textsc{Round}(N \cdot \textsc{PathProb}(\mathbf{v}))$
  $\mathcal{P}_{hit} \leftarrow \textsc{Retrieve}(\mathcal{D}, \mathbf{v}, \text{limit} = n_{tgt})$
  $n_{gap} \leftarrow \max(0, n_{tgt} - |\mathcal{P}_{hit}|)$
  $\mathcal{P}_{miss} \leftarrow \textsc{Augment}(\mathcal{M}, \mathbf{v}, n_{gap})$
  $\mathcal{P} \leftarrow \mathcal{P} \cup \mathcal{P}_{hit} \cup \mathcal{P}_{miss}$
return $\mathcal{P}$

[1] https://www.worldvaluessurvey.org/WVSDocumentationWV7.jsp

by deriving the overall structure from the topic-adaptive tree and preferentially retrieving real data.

3 Benchmark and Evaluation Framework

This section constructs a multi-domain benchmark and proposes PACE, a framework designed to quantify generation quality from statistical alignment to sociological consistency.

3.1 Benchmark Construction

A critical challenge in evaluating population simulation is the absence of the benchmark that explicitly maps specific topics to fine-grained user demographic distributions. Existing corpus typically contain user-generated content, but lack the structured demographic labels necessary for ground-truth comparison. To address this problem, we construct a Topic-based Benchmark by inferring reference populations directly from behavioral data.

We select three publicly available diverse corpus to cover heterogeneous domains: **Bluesky Social Dataset** [2] for social simulation, **Amazon Reviews 2023** [3] for product recommendation, and **IMDB User Reviews** [4] for movie critique. We identify representative topics (e.g., discussion themes, product categories, or movies) and aggregate relevant user posts.

To derive the ground-truth demographics for each topic, we employ a **text-to-persona** pipeline (Ge et al., 2024). Based on the sociological premise that language patterns reflect latent identity, this pipeline infers demographic personas from user texts integrated from social media posts and comments on things using the ability of LLM. This forms a population benchmark under different topics, serving as a reference standard for evaluating the generated population. We also conducted effective manual sampling verification on the constructed population benchmark, showing a consistency rate of 92%. Detailed information on data subsampling, filtering criteria, and topic selection is provided in Appendices A.

3.2 PACE Evaluation Framework

HAG To measure the quality of generated population, we propose **PACE** (**P**opulation **A**lignment & **C**onsistency **E**valuation), which is structured around two aspects: Population Alignment, which quantifies statistical fidelity against ground truth (GT), and Sociological Consistency, which evaluates the sociological rationality of semantics. An overview of the PACE framework is shown in the right part of Figure 2.

3.2.1 Population Alignment

This aspect measures the objective statistical distance between the generated population ($P_{\texttt{Gen}}$) and the real-world ground truth ($P_{\texttt{GT}}$).

Distributional Fidelity. We evaluate the gap between the demographic composition of the generated population and real world with a statistical method. For each attribute dimension k, we use two divergences: the Jensen-Shannon divergence (JSD) and the Kullback-Leibler divergence (KL). The overall fidelity score is defined as the average divergence across all K dimensions:

$$\mathcal{S}_{dist} = \frac{1}{K}\sum_{k=1}^{K} D(P^{(k)}_{\texttt{Gen}} \parallel P^{(k)}_{\texttt{GT}}) \tag{5}$$

where $P^{(k)}_{\texttt{Gen}}$ and $P^{(k)}_{\texttt{GT}}$ denote the marginal probability distributions of the k-th attribute in the generated and ground truth population, respectively. $D(\cdot \parallel \cdot)$ denotes the divergence metric (JSD or KL), and a lower $\mathcal{S}_{dist}$ indicates higher fidelity to match fundamental demographic proportions.

Structural Alignment. In addition to considering marginal distributions, we measure the degree of topic diversity error. We use the Gini-Simpson Index ($GS = 1 - \sum p_i^2$) to quantify diversity, where p_i denotes the proportion of a category i (Simpson, 1949), and define the Diversity Error as the absolute difference between the structural diversity of the generated cohort and the GT:

$$\mathcal{D}_{error} = \frac{1}{K}\sum_{k=1}^{K} |GS^{(k)}_{\texttt{Gen}} - GS^{(k)}_{\texttt{GT}}| \tag{6}$$

A $\mathcal{D}_{error}$ close to zero indicates that the framework adaptively aligns with the topic's social structure, which preserves diverse topics and narrows the focus for niche discussions.

3.2.2 Sociological Consistency

This aspect extends statistics to the sociological semantic quality of the agents.

[2] https://zenodo.org/records/14669616

[3] https://huggingface.co/datasets/McAuley-Lab/Amazon-Reviews-2023

[4] https://www.kaggle.com/datasets/sadmadlad/imdb-user-reviews

Model	Method	Bluesky Social Dataset					Amazon Reviews 2023					IMDB Movies User Reviews				
		Alignment ↓			Consistency ↑		Alignment ↓			Consistency ↑		Alignment ↓			Consistency ↑	
		JSD	KL	DivErr	ArchRel	IndCon	JSD	KL	DivErr	ArchRel	IndCon	JSD	KL	DivErr	ArchRel	IndCon
Data-based Retrieval Methods																
——	Random Select	0.628	2.489	0.505	3.000	2.599	0.530	1.286	0.518	3.000	2.878	0.510	1.359	0.535	2.500	3.440
——	Topic-Retrieval	0.578	5.725	0.285	3.250	2.928	0.587	4.035	**0.408**	3.000	3.134	0.576	2.049	0.473	3.000	3.242
LLM-based Generation Methods																
Average	LLM Generate	0.539	2.487	0.466	3.063	3.197	0.451	0.925	0.504	2.750	**3.675**	0.479	**1.041**	0.555	2.875	3.690
	HAG-Flat	0.401	2.436	0.276	3.750	3.324	0.429	0.779	0.439	3.125	3.487	0.398	3.392	0.324	3.000	3.686
	HAG(Our)	**0.345**	**1.657**	**0.263**	**3.813**	**3.617**	**0.414**	**0.759**	0.419	**3.250**	3.642	**0.393**	1.331	**0.322**	**3.125**	**3.783**
GPT-4	LLM Generate	0.559	1.432	0.541	3.250	2.947	0.515	1.185	0.566	2.000	3.228	0.502	1.176	0.575	3.000	3.565
	HAG-Flat	0.401	1.497	0.295	3.750	3.315	0.421	0.784	0.432	3.000	3.547	0.379	0.914	0.417	3.000	3.785
	HAG(Our)	0.354	1.084	0.281	3.750	3.510	0.447	0.815	0.423	3.000	3.628	0.385	0.734	0.342	3.000	3.828
GPT-oss-120b	LLM Generate	0.485	0.967	0.468	3.250	3.363	0.409	0.688	0.473	3.000	3.685	0.461	0.884	0.521	3.000	3.619
	HAG-Flat	0.411	2.518	0.226	3.250	3.089	0.454	0.895	0.439	3.000	3.543	0.304	0.473	0.326	3.000	3.999
	HAG(Our)	0.320	0.619	0.250	4.000	3.712	0.407	0.754	0.434	3.000	3.643	0.297	0.421	0.272	3.500	4.048
Gemini-2.5-Pro	LLM Generate	0.627	6.513	0.400	2.500	2.911	0.417	0.765	0.442	3.000	4.022	0.481	1.010	0.556	3.000	3.890
	HAG-Flat	0.451	4.665	0.320	3.750	3.317	0.423	0.743	0.412	3.500	3.399	0.434	7.231	0.226	3.000	3.507
	HAG(Our)	0.393	4.052	0.259	3.500	3.558	0.401	0.738	0.431	3.000	3.619	0.423	1.034	0.350	3.000	3.652
DeepSeek-V3.2	LLM Generate	0.485	1.035	0.456	3.250	3.565	0.465	1.061	0.536	3.000	3.765	0.474	1.093	0.568	2.500	3.685
	HAG-Flat	0.343	1.063	0.264	4.250	3.576	0.418	0.695	0.474	3.000	3.457	0.476	4.950	0.326	3.000	3.454
	HAG(Our)	0.313	0.873	0.260	4.000	3.690	0.402	0.727	0.388	4.000	3.676	0.467	3.135	0.323	3.000	3.605

Table 2: Results of experiments across Bluesky, Amazon, and IMDB domains. Metrics for Population Alignment (JSD, KL, DivErr) are *lower-is-better*, while metrics for Sociological Consistency (ArchRel, IndCon) are *higher-is-better*. "Average " denotes the mean score of the four models.

Archetypal Relevance. We evaluate the sociological relevance of the generated "mainstream" voices to the topic. We apply K-Means clustering on the agent embeddings to extract the top-K centroids that keep the dominant archetypes. Then we evaluate the centroids to determine if they represent typical participants in the specific topic.

Individual Consistency. We scrutinize the logical soundness of individual agents through a dual-lens assessment. We conduct a comprehensive evaluation across the entire generated population, ensuring that every generated persona is checked for both internal self-consistency that avoids attribute contradictions, and contextual rationality that ensures plausibility within the specific topic.

4 Experiments

This section reports the experiment setup and results. Further details(including the robustness and sensitivity of the method) are provided in Appendices B.

4.1 Experimental Setup

Evaluation employs a multi-model strategy using LLMs, including GPT-4, GPT-oss-120b, Gemini-2.5-Pro, and DeepSeek-V3.2. Our diverse model selection reduces bias, ensuring findings reflect generalized capabilities. The specific evaluation metrics and baseline methods are detailed below.

Evaluation Metrics. To implement the PACE evaluation framework, we report five quantitative metrics. For Population Alignment, we calculate **JSD** and **KL** to evaluate the distribution fidelity of the population dimension, while using diversity error (**DivErr**) to quantify structural alignment errors. For Sociological Consistency, we score Archetypal Relevance (**ArchRel**) and Individual Consistency (**IndCon**) on a scale of 1-5 points. We use the LLM-as-a-judge method as the paradigm (Li et al., 2024a), and the reliability of this automated judge was also verified through human evaluation.

Baselines. We compare **HAG(Our)** against four baselines categorized into two paradigms, specifically selected for topic-dependent population generation tasks. Within Data-based Retrieval, **Random Select** serves as a topic-agnostic baseline by randomly sampling personas from the WVS pool, while **Topic-Retrieval** selects the Top-N personas based on semantic similarity by embedding the topic and each WVS user's text using a sentence transformer model. Within LLM-based Generation, **LLM Generate** directly generates personas via end-to-end prompting with the topic and persona template. Additionally, we include **HAG-Flat**, an ablation variant that independently generates each demographic dimension distribution (conditioned topic only) to isolate the contribution of our hierarchical dependency modeling.

4.2 Main Results

Table 2 shows the comparison across three domains. Overall, HAG achieves comprehensive improvements in both statistical alignment and semantic consistency. Specifically, compared to the LLM Generate baseline on the Bluesky dataset, HAG obtains an average improvement of **37.7%** on population alignment metrics and **18.8%** on sociological consistency metrics. This superior performance trend is consistent across domains.

Detailed comparisons reveal that Data-based Retrieval methods consistently exhibit poor fidelity and consistency in two evaluation aspects, with Mean JSD scores exceeding 0.500 and IndCon often falling below 3.200 on most domains. Regarding structural alignment, both Random Select and LLM Generate suffer from high errors. We also observed that Gemini-2.5-Pro exhibited significant anomalies on the Bluesky dataset, which inspection traced to world knowledge hallucinations regarding specific topics. Notably, on the IMDB dataset, LLM Generate obtains a high DivErr (0.555), comparable to the noise of Random Select (0.535), but HAG reduces this to 0.322. These results confirm that HAG effectively preserves the complex structural heterogeneity of real-world populations. Furthermore, HAG consistently outperforms the HAG-Flat variant across all domains, validating the effectiveness of our hierarchical structure.

4.3 Qualitative Analysis

Quantitative metrics show HAG's superiority, and we further conduct qualitative analysis to investigate the intrinsic mechanisms. We answer two Research Questions (RQs).

RQ1: Does the population maintain a macro manifold structure consistent with reality? We upgrade our perspective from attribute statistics to the examination of the Population Manifold. We embed personas into high-dimensional dense vectors to capture their complex sociological features, and visualize their local clustering structure in a latent persona embedding space using t-SNE dimensionality reduction. Observed in Figure 3, data-based retrieval methods drift into distal regions, indicating a semantic misalignment with the target population. LLM Generate clusters tightly in a disjoint region, signaling severe mode collapse and a failure to capture the semantic diversity of the topic. Conversely, HAG demonstrates the highest degree of overlap with the GT, effectively filling the target manifold. This confirms that HAG can not only align marginal statistics but also successfully capture the complex, high-dimensional manifold structure of the real-world population.

Figure 3: Visualization of the Population Manifold Structure in the Latent Persona Embedding Space (t-SNE). The area enclosed by the red dashed lines delineates the target mapping region of the Ground Truth population.

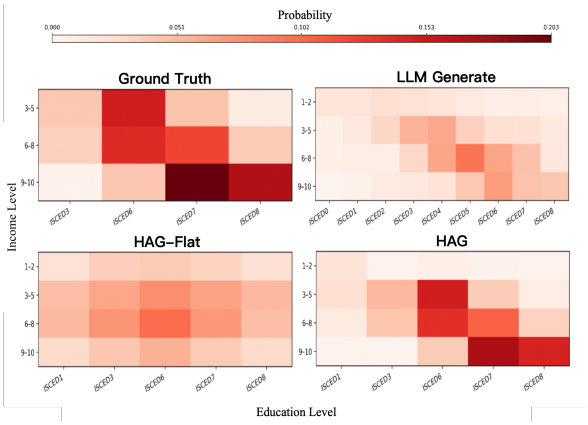

Figure 4: A joint distribution heatmap visualizing the correlation between education level and income level under the scientific research topic, including GT and three LLM-based Generation Methods.

RQ2: Does the population capture joint distributions and conditional dependencies? We visualize the correlation between education level and income level under scientific research topics through joint distribution heatmaps, shown in Figure 4. LLM Generate produces a diffuse spread covering all education levels, including lower levels absent in the GT, indicating a failure to align with the demographic scope. HAG-Flat exhibits a symmetric, grid-like distribution. Because it combines dimension values independently, leading to uniform probabilities. Conversely, HAG closely

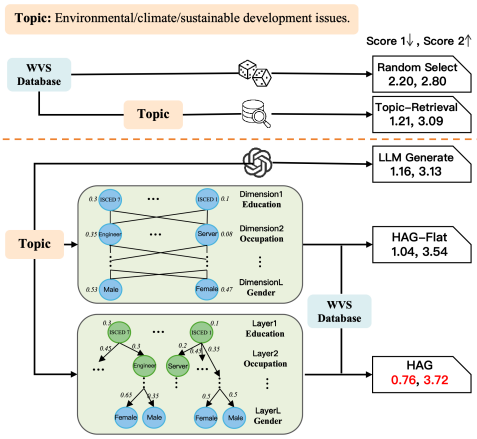

Figure 5: Demonstration of the case study.

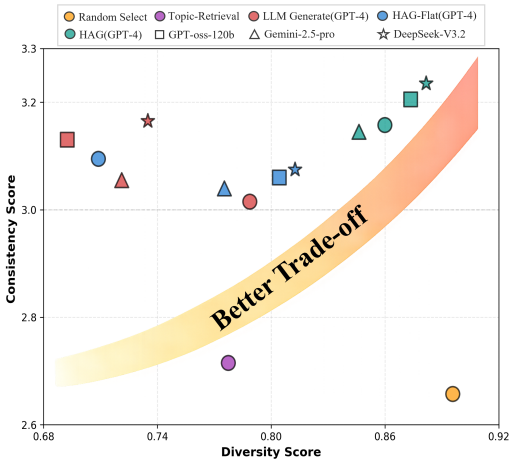

Figure 6: Diversity-Consistency Trade-off on Emerging Topic "Mars Colonization".

matches the diagonal trend and high-density regions of the GT. This validates that our Topic-Adaptive Distribution Tree successfully locks in conditional dependencies, ensuring the generated population conforms to the macro multi-attribute joint distribution, and that the attributes of micro individuals are also relevant and reasonable.

4.4 Case Study

Figure 5 illustrates a case study on the topic of "Environmental". We trace the workflow of five methods and report their average performance on Alignment (Score 1) and Consistency (Score 2). Retrieval methods get high alignment errors. LLM Generate acts as a black box with moderate performance. The HAG-Flat variant constructs personas by combining dimensions independently, which improves consistency but lacks structural precision. HAG explicitly models conditional dependencies via a hierarchical demographic tree, whose structure enables HAG to achieve the best scores.

4.5 Generalization and Adaptability Analysis

To evaluate adaptability on emerging topics (e.g., "Mars Colonization"), we analyze the trade-off between Structural Diversity (Gini-Simpson Index) and Sociological Consistency (defined as $S_{cons} = (\text{ArchRel} + \text{IndCon})/2$).

As illustrated in Figure 6, Random Select occupies the high-diversity but low-consistency region, representing mere statistical noise rather than meaningful heterogeneity. LLM Generate fluctuates in the middle, often succumbing to logical coherence to satisfy diversity. And the performance of LLM Generate and HAG-Flat is easily affected by the models. In contrast, HAG is at the **Pareto Frontier** of this trade-off, maintaining high consistency without compromising population diversity. This confirms that by combining the Topic-Adaptive demographic distribution tree with grounded instantiation, HAG robustly generalizes to open-world scenarios, generating populations that are both diverse and sociologically rational.

5 Related Work

Persona generation. Persona generation turns a population distribution into concrete agent personas, and existing approaches largely follow either retrieval from real-user footprints or LLM-based generation. Retrieval-style pipelines typically build personas from real-world data: some retrieve or enrich existing user personas for a given topic (Yang et al., 2024; Zhang et al., 2025), while others directly infer implicit personas from user logs (Wang et al., 2025a). Generative methods generate personas in two ways: one infers personas from texts by extracting salient information under a certain topic (Ge et al., 2024; Hu et al., 2025); the other generates personas directly under predefined schemas such as persona templates or descriptions (Wang et al., 2025b; Li et al., 2025; Chen et al., 2025; Schuller et al., 2024). However, personas produced by existing methods often fall short in two ways: they either lack a macro-level awareness of the topic or deviate from real-world population distributions, leading to biased persona generation. (Li et al., 2025; Hu et al., 2025)

Social simulation. LLM-driven social simulation replaces simple heuristic behavioral rules with LLM-based policies. In such cases, agent fidelity becomes a key determinant of simulation validity (Wang et al., 2025d). Social media simulators either instantiate scenario-matched subpopulations from real-user pools (Yang et al., 2024; Zhang et al., 2025) or translate scenario descriptions into executable agents and environments (Wang et al., 2025b) to reduce mismatches between intended and implemented behaviors, while recommendation simulators use users' reactions to emulate how platforms decide what to show next (Wang et al., 2025c; Lyu et al., 2024; Cai et al., 2025). Still, structural deviation (e.g., overestimated political homophily) can arise in social simulations, indicating the need for stronger fidelity and calibration checks (Chang et al., 2025; Li et al., 2025).

6 Conclusion and Future Work

We propose HAG, a Hierarchical Demographic Tree-based Agent Generation framework that combines a Topic-Adaptive distribution tree with real-world grounded instantiation, which improves macro-level distribution fidelity and micro-level individual consistency in simulations. Furthermore, we establish a multi-domain benchmark and a comprehensive PACE evaluation framework.

Future work will proceed in two main directions. (1) Benchmark Enrichment: We aim to continuously expand our multi-domain benchmark to encompass more diverse cultural contexts, underrepresented communities, and rapidly emerging social topics. (2) WKM Optimization: We plan to enhance the World Knowledge Model by integrating dynamic Retrieval-Augmented Generation (RAG) driven by real-time sociological databases, which will further minimize temporal lag and mitigate latent model biases.

Limitations

While HAG is robust, its distribution tree construction relies on the WKM's prior knowledge. In rare niche domains, WKMs may exhibit isolated hallucinations (e.g., Gemini-2.5-Pro on specific Bluesky topics). Importantly, these are infrequent edge cases that do not diminish the general efficacy or structural advantages of our hierarchical approach. And as outlined in our future work, we anticipate that integrating dynamic RAG will effectively mitigate this limitation, as grounding the WKM with real-time sociological databases will continuously supply external domain knowledge and overcome the boundaries of static model parameters. Additionally, the instantiation phase utilizes the WVS database. While real-world demographic structures naturally evolve, our framework is highly adaptable. Because the WVS is officially and periodically updated, maintaining the temporal fidelity of the HAG pipeline simply requires synchronizing with the latest data waves, necessitating no structural or algorithmic modifications.

Furthermore, like all LLM-driven pipelines, HAG faces potential "collective bias" stemming from overlapping pre-training paradigms. Although our cross-model validation mitigates model-specific self-alignment, the "social consensus" evaluated by LLMs remains inherently subjective. This systemic challenge extends beyond our specific framework. Ultimately, we hope HAG and PACE serve as transparent baselines for developing more objective, bias-resistant simulation protocols.

Ethics Statement

The ethics statement of this research is as follows.

Data Privacy. Our framework relies exclusively on publicly available datasets. By generating agents from statistical joint distributions rather than raw user logs, HAG serves as a privacy-preserving alternative.

Mitigation of Misuse. We strictly condemn misusing this technology for disinformation or malicious social manipulation. While HAG lowers the barrier for high-fidelity simulation, we urge practitioners to deploy it solely for benevolent, analytical purposes to avoid exploiting vulnerable populations.

Bias and Fairness. Although HAG mitigates statistical biases through empirical grounding, residual biases from underlying LLMs or historical WVS data may persist. For critical applications, we strongly recommend a human-in-the-loop approach to audit generated populations for potential stereotyping before deployment.

Acknowledgments

This work was supported by the Director's Fund Project of State Key Laboratory of AI Safety, the Strategic Priority Research Program of the CAS (No. XDB0680302), and the Young Elite Scientists Sponsorship Program of the Beijing High Innovation Plan(NO.20250924).

References

Ngoc Bui, Hieu Trung Nguyen, Shantanu Kumar, Julian Theodore, Weikang Qiu, Viet Anh Nguyen, and Rex Ying. 2025. Mixture-of-personas language models for population simulation. *arXiv preprint arXiv:2504.05019*.

Shihao Cai, Jizhi Zhang, Keqin Bao, Chongming Gao, Qifan Wang, Fuli Feng, and Xiangnan He. 2025. Agentic feedback loop modeling improves recommendation and user simulation. In *Proceedings of the 48th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 2235–2244.

Serina Chang, Alicja Chaszczewicz, Emma Wang, Maya Josifovska, Emma Pierson, and Jure Leskovec. 2025. Llms generate structurally realistic social networks but overestimate political homophily. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pages 341–371.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, and 1 others. 2024. From persona to personalization: A survey on role-playing language agents. *arXiv preprint arXiv:2404.18231*.

Rongxin Chen, Yunfan Li, Yige Yuan, Bingbing Xu, and Huawei Shen. 2025. Multi-personality generation of llms at decoding-time. *arXiv preprint arXiv:2511.01891*.

Yuanjun Feng, Vivek Choudhary, and Yash Raj Shrestha. 2025. Noise, adaptation, and strategy: Assessing llm fidelity in decision-making. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 7704–7717.

Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. 2024. Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanities and Social Sciences Communications*, 11(1):1–24.

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*.

John J Horton. 2023. Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research.

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*.

Bernard J Jansen, Joni Salminen, Soon-gyo Jung, and Kathleen Guan. 2022. *Data-driven personas*. Springer Nature.

Maik Larooij and Petter Törnberg. 2025. Validation is the central challenge for generative social simulation: a critical review of llms in agent-based modeling. *Artificial Intelligence Review*, 59(1):15.

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

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024a. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

Nian Li, Chen Gao, Mingyu Li, Yong Li, and Qingmin Liao. 2024b. Econagent: large language model-empowered agents for simulating macroeconomic activities. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 15523–15536.

Hanjia Lyu, Song Jiang, Hanqing Zeng, Yinglong Xia, Qifan Wang, Si Zhang, Ren Chen, Chris Leung, Jiajie Tang, and Jiebo Luo. 2024. Llm-rec: Personalized recommendation via prompting large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 583–612.

Emma Rose Madden. 2025. Evaluating the use of large language models as synthetic social agents in social science research. *arXiv preprint arXiv:2509.26080*.

Benjamin S Manning, Kehang Zhu, and John J Horton. 2024. Automated social science: Language models as scientist and subjects. Technical report, National Bureau of Economic Research.

Debaditya Pal and David Traum. 2025. Beyond simple personas: Evaluating llms and relevance models for character-consistent dialogue. In *Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 383–396.

Qiyao Peng, Hongtao Liu, Hua Huang, Qing Yang, and Minglai Shao. 2025. A survey on llm-powered agents for recommender systems. *arXiv preprint arXiv:2502.10050*.

Andreas Schuller, Doris Janssen, Julian Blumenröther, Theresa Maria Probst, Michael Schmidt, and Chandan Kumar. 2024. Generating personas using llms and assessing their viability. In *Extended abstracts of the CHI conference on human factors in computing systems*, pages 1–7.

Edward H Simpson. 1949. Measurement of diversity. *nature*, 163(4148):688–688.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two tales of persona in llms: A survey of role-playing and personalization. *arXiv preprint arXiv:2406.01171*.

Kuang Wang, Xianfei Li, Shenghao Yang, Li Zhou, Feng Jiang, and Haizhou Li. 2025a. Know you first and be you better: Modeling human-like user simulators via implicit profiles. *arXiv preprint arXiv:2502.18968*.

Lei Wang, Heyang Gao, Xiaohe Bo, Xu Chen, and Ji-Rong Wen. 2025b. Yulan-onesim: Towards the next generation of social simulator with large language models. In *Workshop on Scaling Environments for Agents*.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, and 1 others. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345.

Lei Wang, Jingsen Zhang, Hao Yang, Zhi-Yuan Chen, Jiakai Tang, Zeyu Zhang, Xu Chen, Yankai Lin, Hao Sun, Ruihua Song, and 1 others. 2025c. User behavior simulation with large language model-based agents. *ACM Transactions on Information Systems*, 43(2):1–37.

Qian Wang, Jiaying Wu, Zichen Jiang, Zhenheng Tang, Bingqiao Luo, Nuo Chen, Wei Chen, and Bingsheng He. 2025d. Llm-based human simulations have not yet been reliable. *Preprint*, arXiv:2501.08579.

Zixu Wang, Bin Xie, Bingbing Xu, Shengmao Zhu, Yige Yuan, Liang Pang, Long Yang Du Su, Zixuan Li, Huawei Shen, and Xueqi Cheng. A survey on llm-based agents for social simulation: Taxonomy, evaluation and applications.

Yutong Xie, Ruoyi Gao, and Qiaozhu Mei. 2025. Distributional alignment for social simulation with LLMs: A prompt mixture modeling approach. In *First Workshop on Social Simulation with LLMs*.

Ziyi Yang, Zaibin Zhang, Zirui Zheng, Yuxian Jiang, Ziyue Gan, Zhiyu Wang, Zijian Ling, Jinsong Chen, Martz Ma, Bowen Dong, and 1 others. 2024. Oasis: Open agent social interaction simulations with one million agents. *arXiv preprint arXiv:2411.11581*.

Xinnong Zhang, Jiayu Lin, Xinyi Mou, Shiyue Yang, Xiawei Liu, Libo Sun, Hanjia Lyu, Yihang Yang, Weihong Qi, Yue Chen, and 1 others. 2025. Socioverse: A world model for social simulation powered by llm agents and a pool of 10 million real-world users. *arXiv preprint arXiv:2504.10157*.

Wenlong Zhao, Debanjan Mondal, Niket Tandon, Danica Dillion, Kurt Gray, and Yuling Gu. 2024. Worldvaluesbench: A large-scale benchmark dataset for multi-cultural value awareness of language models. *arXiv preprint arXiv:2404.16308*.

Xi Zhu, Yu Wang, Hang Gao, Wujiang Xu, Chen Wang, Zhiwei Liu, Kun Wang, Mingyu Jin, Linsey Pang, Qingsong Weng, and 1 others. 2025. Recommender systems meet large language model agents: A survey. *Foundations and Trends® in Privacy and Security*, 7(4):247–396.

A Evaluation Details

This section details the construction of our evaluation benchmark and the protocols for both automated and human assessment.

A.1 Benchmark Construction

To evaluate population generation in realistic scenarios, we constructed a Topic-Conditioned Demographic Benchmark derived from three datasets: Bluesky Social, Amazon Reviews 2023, and IMDb Movie Reviews.

Data Preprocessing and Topic Selection. For each dataset, we applied rigorous filtering to ensure the quality of the behavioral text used for demographic inference:

- **Data Cleaning:** We filtered out short texts (fewer than 15 tokens) and removed non-natural language content (e.g., spam, URLs). To ensure an appropriate amount of context for profiling, we subsampled the review datasets (Amazon/IMDb) by filtering reviews within a specified time range and retaining users with an appropriate number of posted reviews. For Amazon Reviews 2023, we kept reviews posted in 2023 and retained users with at least 10 reviews. For IMDb Movie Reviews, we kept reviews posted in 2021–2022.
- **Topic Selection:** We selected topics based on two criteria: (1) *Volume*, ensuring at least 50 unique users participated in the discussion; and (2) *Controversy/Diversity*, ensuring the topic elicited a wide range of sentiments and perspectives. Guided by these criteria, we selected eleven discussion themes in Bluesky, two product categories in Amazon, and two movies in IMDb as our topics. The selected topics are detailed in Table 3.

Text-to-Persona Pipeline. Since public datasets lack fine-grained demographic labels, we employed a Text-to-Persona approach to infer Ground Truth personas from user-generated content. The attributes of personas are derived from the demographic attributes extracted from the WVS dataset, as shown in Table 1. We utilized **GPT-4o** as the inference engine due to its superior reasoning capabilities. The model accepts a user's historical posts/reviews as input and infers demographic attributes (Age, Gender, Education, Income, Occupation, Social Class). To minimize hallucinations,

Dataset	Theme/Category/Movie	Topic
Bluesky Social Dataset	#Disability	People with disabilities and disability issues.
	#UkrainianView	Ukrainian perspectives and experiences during the war.
	AcademicSky	Academic discussions, including higher education, academic work, academic discourse, etc. (mainly aimed at academic groups).
	BlackSky	The voices and issues of black users (the group consists of individuals who identify themselves as black users).
	BookSky	Reading, book recommendations, literature.
	Game Dev	Game development topics, including production, programming, design, etc..
	GreenSky	Environmental/climate/sustainable development issues (such as climate change, emissions, energy, etc.).
	News	Headline content released by news organizations (which may include major current events in various countries).
	Political Science	Research and Discussion in the Field of Political Science/International Relations.
	Science	Science communication, academic/research personnel, scientific topics and popular science content.
	What's History	Historians/Historical Topics: Historical Research, Historical Stories, Historical Figures.
Amazon Reviews 2023	Baby products	Baby products such as diapers, feeding supplies, and baby care essentials.
	Musical Instruments	Musical instruments and related gear such as guitars, keyboards, drums, and recording equipment.
IMDB Movies User Reviews	Forrest Gump	Forrest Gump (1994)
	Joker	Joker (2019)

Table 3: Topics selected from each dataset.

the model was instructed to return "Unknown" if the text provided insufficient clues.

A.2 Automated Evaluation Setup

For the *Sociological Consistency* metric within our PACE framework, we adopted the **LLM-as-a-Judge** paradigm. We employed **GPT-4o** as the evaluator to score generated agents on two dimensions: Archetypal Relevance (ArchRel) and Individual Consistency (IndCon). The judge evaluates agents on a 5-point Likert scale, providing reasoning for each score.

A.3 Human Verification Protocol

To ensure the validity of our evaluation pipeline, we conducted a rigorous human verification process. This process served two purposes: (1) Benchmark Validation: Verifying the accuracy of the GPT-4o inferenced personas (Ground Truth). (2) Judge Reliability: Verifying the alignment between the automated LLM judge's scores and human sociological assessment.

Expert Annotators. We recruited 10 PhD candidates in Sociology from universities to serve as expert annotators. All annotators possessed advanced knowledge of quantitative research methods and social stratification. They were given the same detailed instructions as the prompts we used when evaluating via LLM-as-a-judge to ensure consistency. Participants were financially compensated for their time, with a payment rate determined in accordance with the standard stipend for graduate research assistants.

Adaptive Sampling Strategy. In both benchmark construction and experimental generation, the total population size M varies significantly across different topics (e.g., niche discussions may contain fewer than 100 users, while broad social topics involve thousands). A fixed sample size would thus be statistically invalid: it would either undersample large populations or oversample small ones. To address this, we employed an Adaptive Random Sampling strategy based on finite population correction.

The sample size n is calculated as follows:

$$n = \min\left(M, \max\left(30, \frac{n_0}{1 + \frac{n_0 - 1}{M}}\right)\right) \quad (7)$$

$$n_0 = \left(\frac{Z \cdot \sigma}{E}\right)^2 \quad (8)$$

where M is the total population size for a specific

topic, n_0 is the initial sample size for an infinite population, Z is the Z-score corresponding to the confidence level (1.96 for 95% confidence), σ is the estimated standard deviation (set to 1.0 as a conservative estimate for Likert-scale variance), and E is the acceptable margin of error (set to ± 0.2). This formula first calculates the initial sample size based on the infinite population assumption, then adjusts it using the finite population correction factor, ensuring that the final sample size is neither less than 30 (to maintain statistical reliability) nor exceeds the total population size M.

The specific sampling statistics derived from this protocol are detailed in Table 4.

Pop. (M)	**Sample** (n)	**Prop. (%)**	**Error** (E)
20	20	100.00	0.00
30	30	100.00	0.00
31	30	96.77	0.07
100	50	50.00	0.20
101	50	49.50	0.20
500	81	16.20	0.20
501	81	16.17	0.20
1000	88	8.80	0.20
1001	88	8.79	0.20
2000	92	4.60	0.20

Table 4: Adaptive Sampling Protocol Statistics.

Results from this verification showed a **92%** agreement rate between the expert annotators and the GPT-4o derived benchmark/scores, validating the reliability of our automated pipeline.

Internal validation protocols. Beyond our human-concordance check, we implemented internal validation protocols:

- **Signal-Dependency Test (Masking):** To test if the LLM hallucinates unobserved traits, we sampled user texts containing explicit demographic declarations (e.g., stated Income Level or Religion). We then ran the extraction on a masked version of the text where these specific signals were artificially removed. The pipeline correctly abstained (outputting "Unknown" rather than hallucinating a guess) in 94.3% of the masked cases. This proves the extraction is strictly contingent on the presence of textual evidence, not on latent LLM priors.
- **Repeatability Test (Stability):** To ensure the model wasn't relying on unstable, ad-hoc heuristics, we ran the inference pipeline three independent times on the same set of user texts. The inferred personas demonstrated a 95.8% consistency rate across the three trials.

B Experimental Details

This appendix provides additional details regarding the experimental setup, including embedding model selection and parameter settings for our proposed HAG framework.

B.1 Embedding Model

For retrieval and clustering tasks in baselines and evaluation stage, we used the `sentence-transformers/all-MiniLM-L6-v2` [5] model to encode the input topic and all personas from the processed WVS dataset and the generated population into dense embeddings. We set the batch size to 1 when encoding the topic and to 32 otherwise. These embeddings were used to compute cosine similarity between the topic and user personas in the Topic-Retrieval baseline and to perform K-means clustering for the Archetypal Relevance metric.

B.2 Method Parameters

Our proposed HAG framework sets specific parameters for constructing the Topic-Adaptive demographic distribution tree. These parameters were fixed across our experiments as follows:

- Maximum number of prioritized dimensions (max_depth): 5
- Maximum number of values per dimension ($max\_branches$): 5
- Sampling temperature of the world knowledge model ($temperature$): 0.0

B.3 Robustness to Data Sparsity: HIT vs. MISS Analysis

To evaluate the robustness of our framework against data sparsity, particularly in highly niche domains, we analyze the HIT (successful matching with empirical World Values Survey users) versus MISS (reliance on agentic augmentation) ratios during the instantiation phase. Table 5 details these statistics across four distinct themes from our benchmark: BlackSky, Game Dev, AcademicSky, and GreenSky.

[5] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

The statistical distribution reveals that even for highly specialized or niche topics, the HIT rates consistently dominate the generation process (ranging from 64.74% to 68.65%). Notably, there is no significant numerical degradation in the HIT ratio for these niche communities compared to more general or popular topics. This empirical evidence demonstrates that our hierarchical conditional system successfully handles data sparsity without over-relying on the LLM-driven augmentation step, thereby maintaining strong empirical sociological grounding across diverse simulation scenarios.

Topic	HIT Ratio	MISS Ratio
BlackSky	68.37%	31.63%
Game Dev	68.65%	31.35%
AcademicSky	68.47%	31.53%
GreenSky	64.74%	35.26%

Table 5: HIT and MISS ratios across different topics.

B.4 Sensitivity to the Distribution Tree Structure

Sensitivity to Tree Depth. We evaluated varying tree depths (3, 6, and 9) and observed a clear trade-off curve across both macro and micro metrics, as summarized in Table 6. A shallow tree (depth=3) lacks the expressiveness to capture complex joint distributions, resulting in the poorest macro-distribution Alignment (0.875, where lower is better) and the lowest micro-level Consistency (3.56), despite having a high initial HIT Ratio (79.49%). The appropriate setting (depth=6) strikes the best balance, achieving the strongest Alignment (0.5041) and the highest Consistency (3.71) while maintaining a healthy WVS HIT Ratio (66.74%). Conversely, an overly deep tree (depth=9) imposes excessively strict constraints. As the conditional path grows longer, the database struggles to find exact matches (data sparsity), causing the HIT Ratio to plummet to 24.55%. This forces the system to rely heavily on "Agentic Augmentation," triggering error propagation that degrades both Alignment (0.5822) and Consistency (3.62).

Depth	HIT Ratio	Alignment (↓)	Consistency (↑)
3	79.49%	1.575	3.56
6	66.74%	**0.5041**	**3.71**
9	24.55%	0.5822	3.62

Table 6: Performance metrics across varying tree depths.

Sensitivity to Dimension Selection. To evaluate the selection of dimensions, we compared our HAG method (Topic-Adaptive Tree) against a baseline Fixed-Dimension Tree. In the HAG, the WKM dynamically selects dimensions and priorities. For example, for the AcademicSky topic in the Bluesky benchmark, the WKM dynamically selected **Education** → **Occupation** → **Country** → **Language** → **Age**. Conversely, the Fixed-Dimension Tree enforces the use of a fixed and universal set of sociological dimensions: **Age** → **Gender** → **Country** → **Education** → **Income**. As shown in Table 7, the adaptive HAG tree significantly outperforms the fixed tree in terms of alignment (0.573 vs. 1.205) and consistency (3.63 vs. 3.50). This proves that fixed trees waste hierarchical capacity on topic-irrelevant attributes while omitting key features, confirming that dynamic dimension selection is crucial for maintaining topic adaptability.

Method	Alignment (↓)	Consistency (↑)
Fixed-Dimension	1.205	3.50
HAG (Topic-Adaptive)	**0.573**	**3.63**

Table 7: Comparison of dynamic vs. fixed dimension selection.

In summary, these empirical results confirm our intuition: a carefully balanced tree depth is critical to maximize joint distribution expressiveness without triggering data sparsity and error propagation, while dynamic dimension selection is vital to capture topic-specific sociological features. Together, these findings strongly validate the necessity of our WKM-driven, topic-adaptive tree construction approach.

C Prompts

Prompt for identifying prioritized topic-relevant dimensions

```
You are a computational sociologist. Your task is to determine the most important user profile dimensions for a social network simulation on the topic "{topic}".

Please identify up to 5 most critical demographic dimensions from the table below and rank them in descending order of their influence on people's opinions and behaviors related to this topic.

The dimensions in the table are:
1. Gender
2. Age
3. Education
4. Country
5. Language
6. Marital status
7. Occupation
8. Financial status
9. Social class
10. Income level
11. Religion
12. Ethnicity

Please output a list in the following JSON format strictly:
{{
"dimensions": ["Dimension 1", "Dimension 2", "Dimension 3", "Dimension 4"]
}}
```

Prompt for generating a conditioned distribution over a demographic dimension

```
You are a computational sociologist. {context_str}, please generate a plausible probability distribution for the dimension "{dimension}".

List the primary values for this dimension and assign a probability to each.

Provide the most relevant and meaningful values for this context - you can provide anywhere from 1 to {max_branches} values, depending on what makes sense for the given context.

IMPORTANT: Choose the number of values based on what is actually meaningful and significant for this specific context.
- If only 1-3 categories are truly relevant, use only 1-3 values.
- If more categories are meaningful, you can use up to {max_branches} values.
- Do NOT artificially inflate the number of categories just to reach the maximum.

Focus on the most significant categories rather than trying to fill up to the maximum number.
The sum of all probabilities must be exactly 1.0.

Strictly adhere to the following JSON format for your output:
{{
"distribution": [
{{"value": "Value 1", "probability": 0.xx}},
...
]
}}


{allowed_clause}
```

Prompt for Text-to-Persona Pipeline

```
You are a computational sociologist analyzing social media posts to generate realistic user profiles.

TASK: Generate a user profile based on the provided social media posts from the "{theme}" community.
{dimension_info}
USER'S POSTS:
{user_text}

INSTRUCTIONS:
1. Analyze the user's posts to infer their demographic characteristics. Make reasonable inferences
based on the content, language, and context of the posts.
2. Generate a realistic user profile that matches the template structure. Only generate values for
the dimensions specified in the template.
3. Replace all "__FILL__" placeholders with appropriate values. Ensure all values are chosen from
the allowed constraints below.

ALLOWED VALUES (choose ONLY from these lists):
{constraints_text}

IMPORTANT CONSTRAINTS:
- You MUST choose values ONLY from the allowed lists above. Do NOT invent new values or categories.
- If you cannot determine a value from the posts, choose the most common/general option from the
allowed list.
- Consider the theme context: "{theme}" community members may have specific characteristics.
- Only generate the dimensions specified in the template - do not add extra fields.

OUTPUT FORMAT:
Return a JSON object that exactly matches this template structure:
{template_json}

ANALYSIS GUIDELINES:
- Age: Infer from language style, references to life events, generational markers.
- Education: Consider vocabulary, topic complexity, academic references.
- Country: Look for location mentions, cultural references, language patterns.
- Occupation: Analyze professional topics, work-related discussions.
- Religion: Consider religious references, holidays, cultural practices.
- Other fields: Make reasonable inferences based on available information.

Generate the user profile now:
```

Evaluation prompt for Archetype Relevance

```
You are an expert computational sociologist.

DOMINANT CLUSTERS:
{dom_snippet}

THEME / TOPIC CONTEXT:
{theme_context}

Question:
Are these dominant archetypes (typical groups) the core stakeholders for this topic?
Consider whether age, education, occupation, country, language and other demographics
form plausible and meaningful typical user types that align with sociological expectations.

Scoring guide:
- 1: Archetypes are completely irrelevant or implausible for this topic
- 3: Archetypes are somewhat relevant but have some issues
- 5: Archetypes are highly relevant and plausible as core stakeholders for this topic

Return format (must be a valid JSON object):
{{
"archetype_coherence_score": <int 1-5>,
"reasoning": "<short explanation>"
}}
```

Evaluation prompt for Individual Consistency

```
You are a computational social scientist who studies population structure.

Given the following theme/topic and a single agent profile, evaluate whether the
combination of demographic attributes (such as age, education, occupation, country, etc.)
is internally consistent and realistic for that theme.

Topic:
{context}

Agent Profile (JSON):
{user_profile}

Your task:
- Please only judge from the perspective of “logical consistency” and give a rating of 1–5 to
indicate whether the attribute combination of the agent is reasonable and consistent in this
real-world topic:
1 = Very unreasonable (with obvious contradictions)
3 = Generally reasonable (with some doubts but acceptable)
5 = Very reasonable (there is no obvious contradiction)

Return format (must be a valid JSON object):
{{
“internal_consistency_score”: <int 1–5>,
“reasoning”: “<Short text explaining the main reason for judgment>”
}}
```