Paper ID: 2920

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

Rongxin Chen[1,2], Tianyu Wu[1,2], Bingbing Xu[1], Jiatang Luo[1,3], Xiucheng Xu[1,2], Huawei Shen[1,2]

[1]State Key Laboratory of AI Safety, Institute of Computing Technology, CAS

[2]University of Chinese Academy of Sciences

[3]School of Advanced Interdisciplinary Sciences, University of Chinese Academy of Sciences

Code Paper WeChat

Presenter: Rongxin Chen

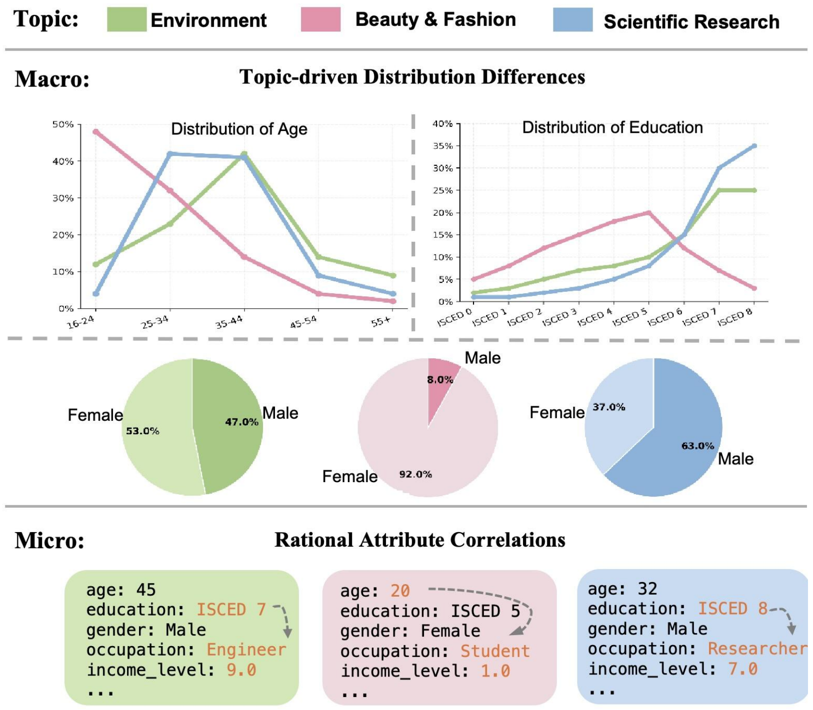
Email: chenrongxin24s@ict.ac.cn

- **Motivation**
- **Method**
- **Experiments**

MOTIVATION

Agent Generation for Simulation

- Agent generation for simulation requires the generated population to be **topic-adaptive**, with macro-level **distribution alignment** and micro-level **persona consistency** across different topics.



MOTIVATION

Existing work

- Data-based Retrieval
  - Construct agent populations by retrieving real-world user profiles or logs from historical data
  - Static and data-coupled, making it hard to generalize to unseen or data-scarce topics.

Real-world User Logs / Historical Data
---Retriever / Matching---> Agent Population

MOTIVATION

Existing work

- LLM-based Generation
  - Construct agent populations by generating independent personas from schemas or text prompts
  - Lack explicit joint distribution modeling, resulting in weak macro-level alignment and possible micro-level inconsistency.

Schema / Text Prompt ---LLM Generator--->
Independent Personas ---> Agent Population

MOTIVATION

Main Idea

- **Observation**: Existing agent generation methods fail to jointly ensure topic-adaptive macro-level population modeling and micro-level persona rationality.
- **Main Idea**: Instead of retrieving static populations or generating independent personas, can we generate a topic-adaptive population **through a hierarchical demographic tree and grounded instantiation**, so that the population is aligned at the macro level and coherent at the micro level?

- Motivation
- **Method**
- Experiments

PROBLEM DEFINITION

Agent Generation for Simulation

- **Goal:** We want to construct a topic-based agentic population dataset grounded in real demographic data.
- **Given:** Topic: t

  Target population size: N
- **Target:** Generate an agent population $\mathcal{P}$ whose demographic composition $\widehat{\mathbf{D}_{\mathcal{P}}}$ matches the target demographic structure $\mathbf{D}(t)$ of that topic.

METHOD: HAG Framework

Hierarchical Agent Generation Framework

- **Two Stages**
  1. Constructing a **hierarchical** demographic distribution tree to capture **topic-relevant** demographic dimensions and their **joint probabilities** of values.

  2. Grounding the tree in real-world data by **filtering** real users and **augmenting** generated agents where data is missing.

METHOD: HAG Framework

Hierarchical Agent Generation Framework

- **Stage 1: Topic-Adaptive Demographic Distribution Tree Construction**
  1. Prioritize: Use a World Knowledge Model to prioritize a subset $\mathbf{f}$ of relevant demographic **dimensions** from a pre-specified set.

  2. Construct: Build the tree top-down along the ordered sequence $\mathbf{f} = \left(f^{(1)}, \ldots, f^{(\mathrm{L})}\right)$.
     a. **Layer** l ($l = 1, 2, \ldots, \mathrm{L}$): the l-th **dimension** $f^{(l)}$ in $\mathbf{f}$
     b. **Nodes** in Layer l : **values** $v^{(l)}$ of the l-th dimension $f^{(l)}$
     c. **Weighted Edges** between Nodes in Layer $l-1$ and Layer l :
        conditional probabilities $w(v^{(l)} \mid v^{(1:l-1)}, t) = P(f^{(l)} = v^{(l)} \mid f^{(1:l-1)} = v^{(1:l-1)}, t)$

  - **Root Node: topic**
  - **Leaf Node with Path Weight:**
    a complete demographic persona $\boldsymbol{v} = (v^{(1)}, \ldots, v^{(L)})$ **with proportion** $W(\boldsymbol{v}|t) = \prod_{l=1}^{L} w\left(v^{(l)}|v^{(1:l-1)}, t\right)$

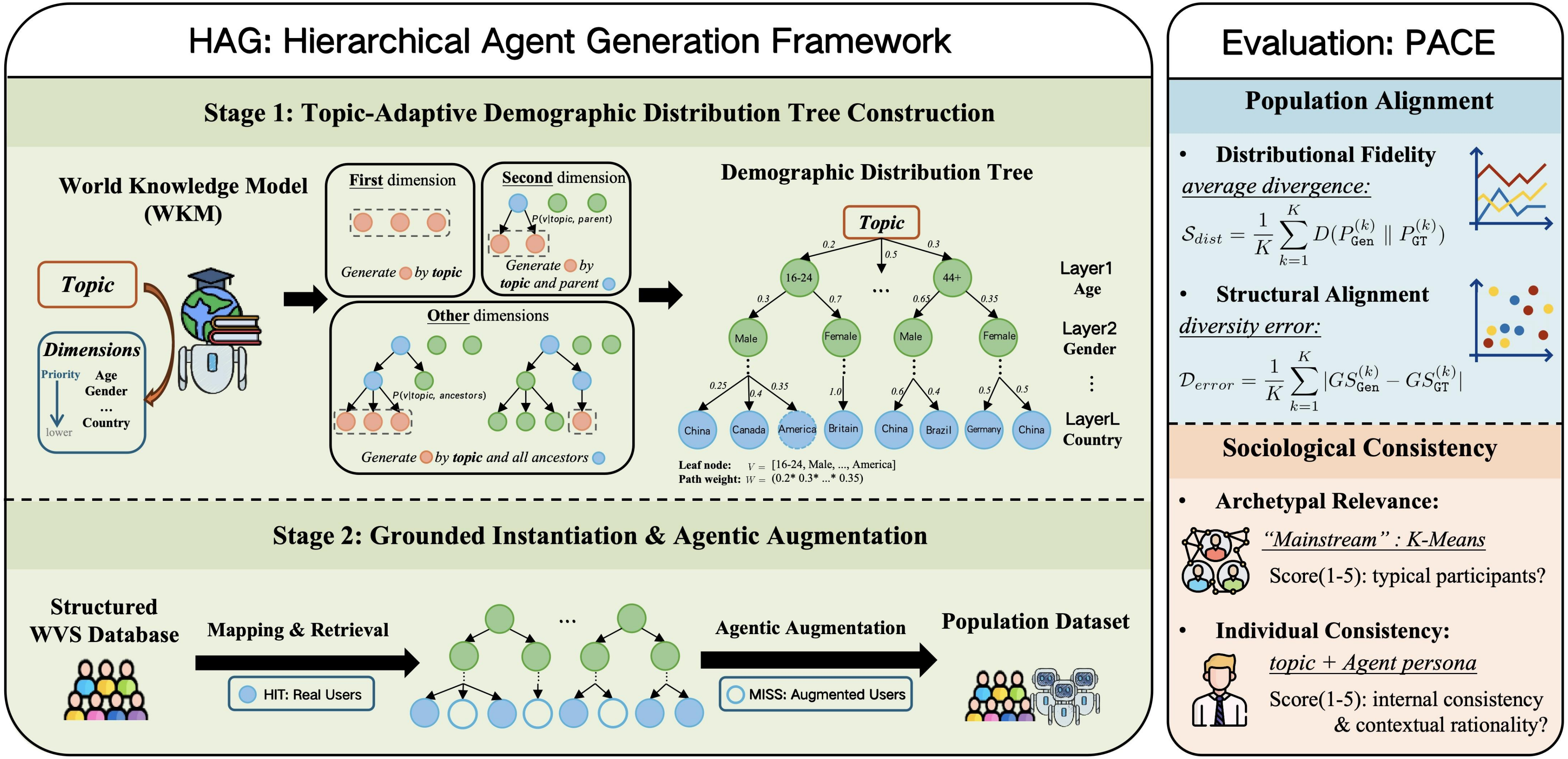
METHOD: HAG Framework

Hierarchical Agent Generation Framework

- **Stage 2: Grounded Instantiation & Agentic Augmentation**
  1. Mapping & Retrieval: **retrieve** real users with persona $\boldsymbol{v}$ from processed WVS database, and **sample** up to $N \times W(\boldsymbol{v}|t)$ users.
  2. Agentic Augmentation: when the retrieved users are insufficient, apply agentic augmentation to fill the vacancy.

METHOD: HAG Framework

Hierarchical Agent Generation Framework



- Motivation
- Method
- **Experiments**

EXPERIMENTS

Experimental Settings

- **Datasets**
  - Bluesky Social Dataset: topics of social simulation
  - Amazon Reviews 2023: topics of product recommendation
  - IMDB User Reviews: topics of movie critique
- **Evaluation: P**opulation **A**lignment & **C**onsistency **E**valuation (PACE)
  - **LLM-as-a-Judge**
  - **Population Alignment:** statistical fidelity against ground truth
    - **Metrics:** JSD, KL, Diversity Error (DivErr)
  - **Sociological Consistency:** the sociological rationality of semantics
    - **Metrics:** Archetypal Relevance (ArchRel), Individual Consistency (IndCon)

EXPERIMENTS

Experimental Settings

- **Baselines**
  - **Random Select:** randomly samples personas from the WVS pool.
  - **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.
  - **LLM Generate:** directly generates personas via end-to-end prompting with the topic and persona template
  - **HAG-Flat:** an ablation variant that independently generates each demographic dimension distribution (conditioned topic only) to isolate the contribution of our hierarchical dependency modeling.

EXPERIMENTS

Main results

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: Results of experiments across Bluesky, Amazon, and IMDB domains.
↑ indicates higher is better, while ↓ indicates lower is better.

EXPERIMENTS

Analysis 1: macro manifold structure

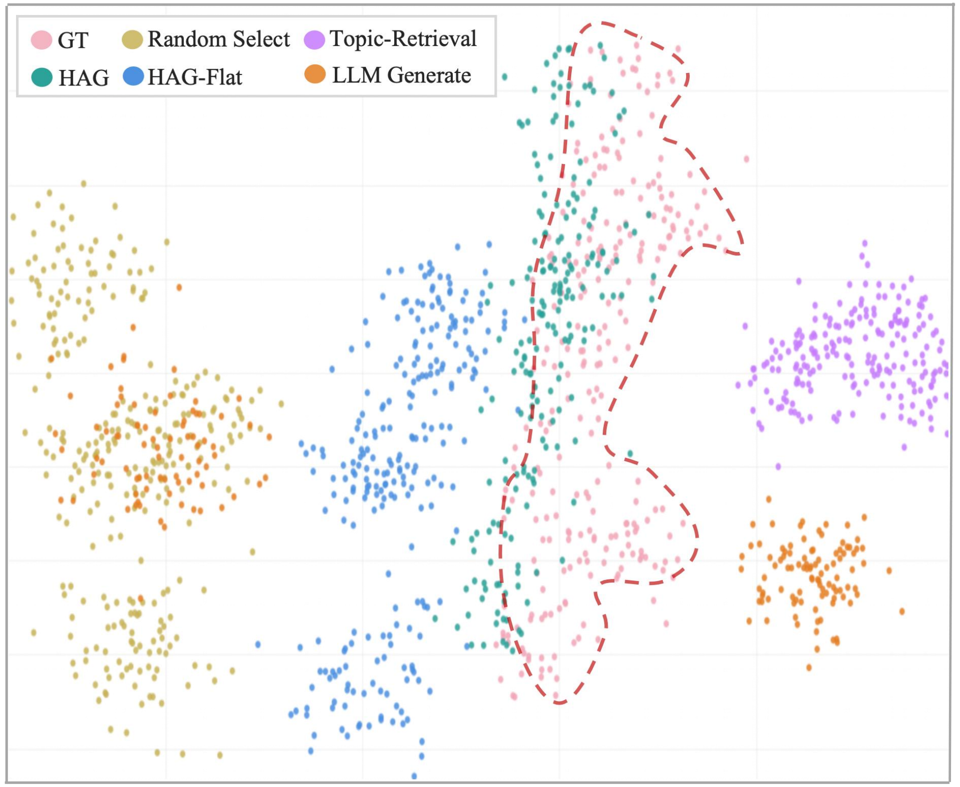


Figure: Visualization of the Population Manifold Structure in the Latent Persona Embedding Space (t-SNE).

EXPERIMENTS

Analysis 2: joint distributions and conditional dependencies

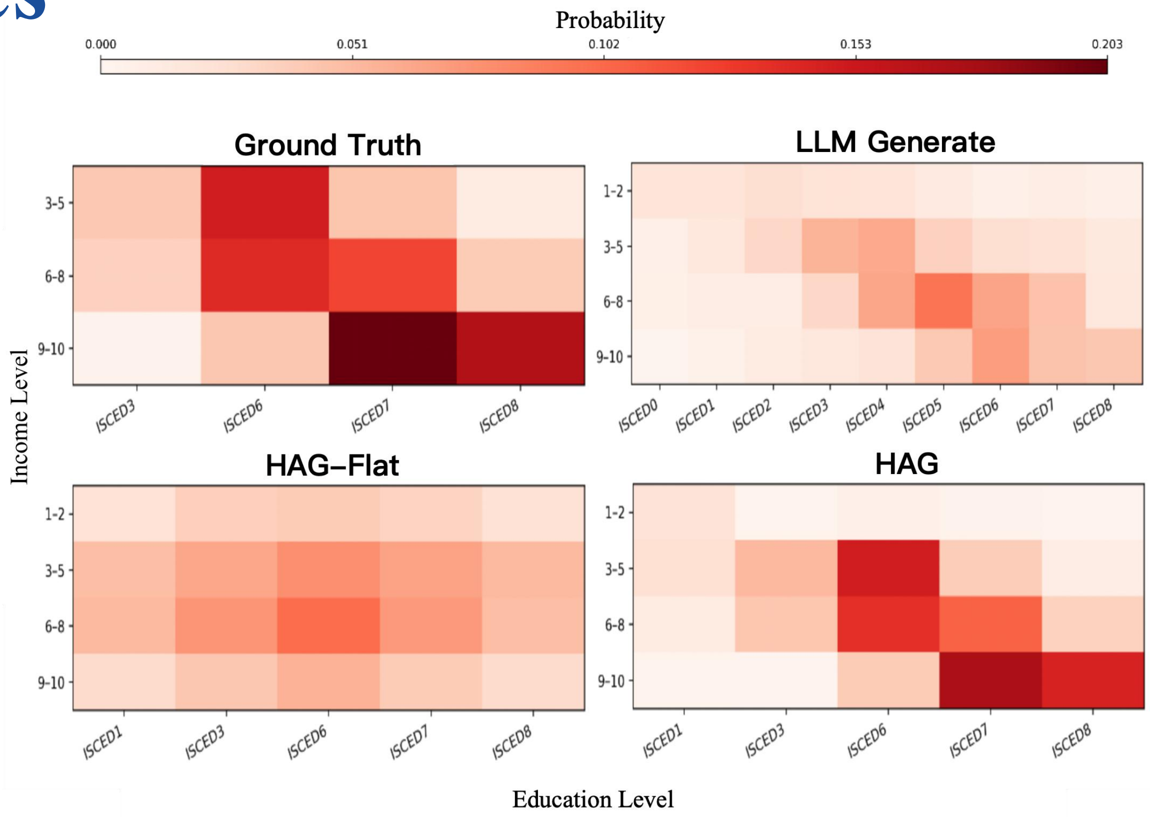


Figure: A joint distribution heatmap visualizing the correlation between **education level** and **income level** under the **scientific research** topic.

EXPERIMENTS

Analysis 3: generalization and adaptability

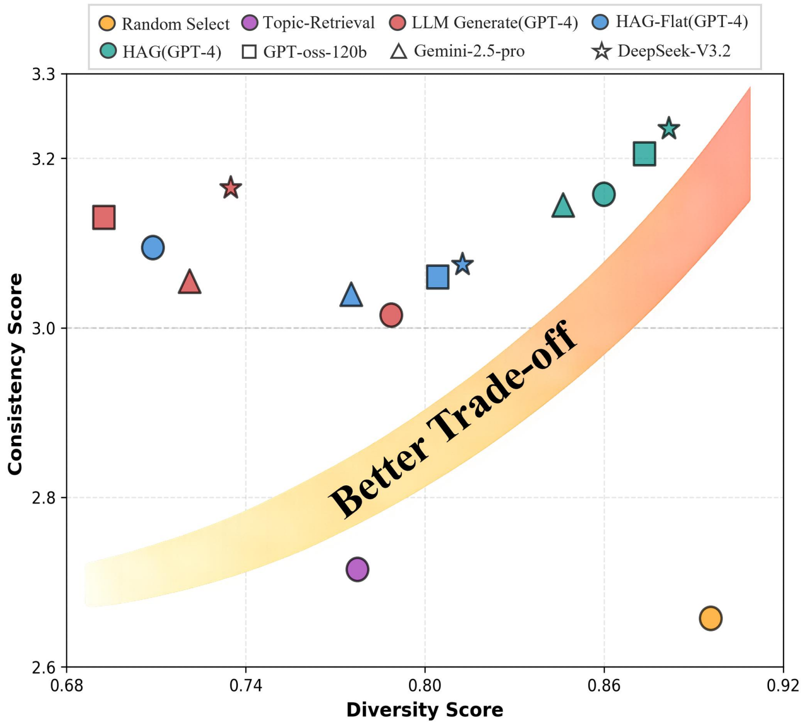
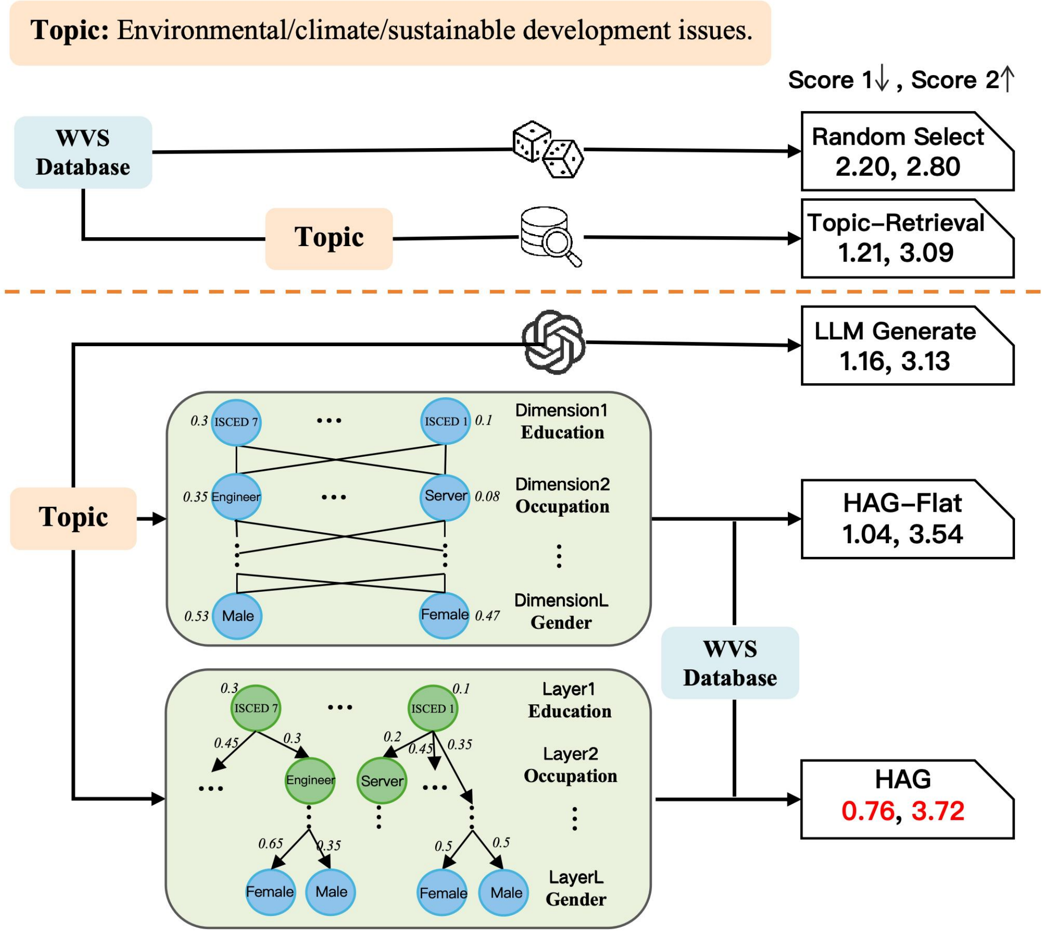


Table: Diversity-Consistency Trade-off on Emerging Topic "Mars Colonization".

HAG robustly generalizes to open-world scenarios, generating populations that are both diverse and sociologically rational.

EXPERIMENTS

Case study



Topic: Environmental/climate/sustainable development issues.

Figure: Demonstration of the case study.
Score 1 indicates average performance on Alignment, and Score 2 indicates average performance on Consistency.

Thank you for your attention!

Code

Paper

WeChat

Email: chenrongxin24s@ict.ac.cn