UNLOCKING MANY VOICES IN LLM SOCIAL SIMULATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Constructing populations for LLM social simulation requires agents that reflect differences among real people. Prompting many agents from a shared model often produces a small set of similar voices, compressing the diversity of the resulting society. To address this problem, we propose ReMap, a method that learns individual variation from real writing and composes it into synthetic populations. We first encode individual variation into model weights through lightweight adaptation. We then compose these learned adaptations across model depth to create agents assembled from multiple source individuals. Experiments across three domains and six backbones show that ReMap populations stay close to human writing while supplying far more mutually distinguishable agents than the library contains, and that they produce more diverse collective behavior than prompted populations inside external simulators. These results suggest a path toward more representative LLM populations for studying collective behavior. Our code and data are anonymously available at `https://bit.ly/4iHRpBb`.

1 INTRODUCTION

Multi-agent simulation powered by large language models has emerged as a methodology for computational social science (Aher et al., 2023; Hua et al., 2023). By constructing artificial societies, researchers can study public opinion dynamics and evaluate potential interventions before deployment (Gao et al., 2023; Piao et al., 2025). The validity of these simulations ultimately depends on the populations that inhabit them. Capturing human heterogeneity is essential to interpreting their collective behavior (Argyle et al., 2023; Santurkar et al., 2023; Bisbee et al., 2024).

However, platforms such as Generative Agents (Park et al., 2023), OASIS (Yang et al., 2024), and AgentSociety (Piao et al., 2025) typically differentiate agents through personas supplied to a shared LLM. Multiplying personas can still produce similar styles and opinions because the agents share the same generation model (Chen et al., 2026; Wu et al., 2025; Ye et al., 2026). Prompt-based sampling seeks additional response modes (Zhang et al., 2025), yet those modes remain responses elicited from the shared model. Population reweighting aligns selected profiles with target marginals (Hu et al., 2025), leaving variation outside those marginals unspecified. Attribute injection supplies demographic differences (Wang et al., 2026), while its population remains organized around the attributes selected for injection.

Human variation extends beyond the dimensions researchers can specify in a profile (Hu & Collier, 2024). Individuals differ in their language and reasoning as well as their beliefs and behavioral tendencies. For example, people with the same stated opinion may differ in what evidence they accept or when they stop arguing. A profile that assigns the opinion leaves these differences unspecified. Providing writing histories gives the model access to such evidence, but the resulting behavior still depends on how the shared model uses that context (Lutz et al., 2025). The challenge is to *recover observed variation without requiring its dimensions to be enumerated in advance*.

To address this, we propose **Re**combining **M**odular **A**daptations into **P**opulations (ReMap), which constructs synthetic populations by recombining adaptations learned from individual writing. Person-specific Low-Rank Adaptation (LoRA) modules encode observed variation on a shared language model. Each agent draws depth blocks from the modules of multiple source people and retains the resulting combination during simulation. A finite module library therefore supports a larger population of composition recipes. Figure 1 illustrates the variation that motivates this construction.

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

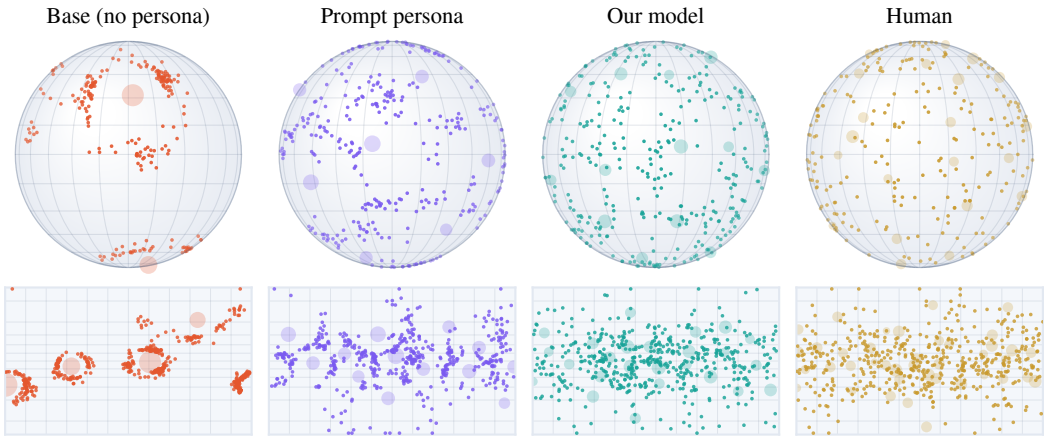

Figure 1. Text variation in base, persona, adapted, and human populations. Positions show the geometry of embedded posts.

We conduct experiments across *Reddit discussion*, *Supreme Court oral arguments*, and *congressional social media* using six LLM backbones. Our ReMap populations move closer to human text distributions and generate greater population diversity than the shared base models. ReMap also expands the library into ten times as many agents that remain distinguishable by their responses, whereas reusing complete modules is bounded by the number of source people. The population comparison also transfers to **OASIS** and **YSocial**. In their opinion and misinformation tasks, changing the population changes disagreement and minority responses, linking text-level variation to behavior measured by a social simulation.

Our contributions concern population construction and its consequences for social simulation. **First**, we identify diversity collapse in persona-based populations and link it to using predefined profiles. **Second**, we introduce ReMap, a cross-layer composition framework that constructs agent populations from adaptations learned on real writing. **Third**, our experiments show that ReMap supplies hundreds of distinguishable agents from forty source modules while keeping the population close to human text, and that this construction changes collective responses inside external simulators.

2 Related Work

Social Simulation. LLM simulation platforms combine agent memory and interaction rules with persona-based population construction (Park et al., 2023; Yang et al., 2024; Piao et al., 2025). Studies of collective generation identify contraction of response diversity during interaction (Chen et al., 2026; Patel, 2026), motivating explicit evaluation of population behavior (Wu et al., 2025; Ye et al., 2026). Prompt sampling (Zhang et al., 2025; Lanchantin et al., 2025), population alignment (Hu et al., 2025), and identity injection (Wang et al., 2026) address different aspects of this variation. Every construction in this line differentiates agents through text supplied to one shared model, and the writing its population produces is left uncompared with human writing.

Population Modeling. Silicon sampling evaluates whether conditioned models reproduce human opinion distributions (Sun et al., 2024; Dominguez-Olmedo et al., 2024). Training on survey responses extends this approach to learned population distributions (Cao et al., 2025; Suh et al., 2025); the output is a distribution over answers with no individual agent behind it. Interview-based agents model individual answers from personal evidence (Park et al., 2024), and personalized adapters and character models fit one model or module to one user for that user's own tasks (Tan et al., 2024b;a; Liu et al., 2024; Shao et al., 2023; Thakur et al., 2025; Guo et al., 2026). These constructions stop at the observed people, and the population their agents would form is never the object of evaluation. Social simulation needs a population that matches real variation and contains more agents than the people it was learned from.

Adapter Composition. Low-rank adaptation represents model updates with compact factors (Hu et al., 2022), and multi-adapter serving shares a backbone across requests (Sheng et al., 2023). Weight arithmetic combines trained changes (Ilharco et al., 2023; Wortsman et al., 2022; Prabhakar

et al., 2025), including layer-wise author adaptation for style transfer (Thillainathan et al., 2026). Adapter populations have also been studied through reasoning self-play (Creus Castanyer et al., 2026) and personality traits (Baines et al., 2026). Our evaluation connects cross-layer resampling with human text distributions and the collective responses of the resulting population.

3 METHOD

3.1 PROBLEM DEFINITION

Let f_{θ_0} be a shared language model, and let K observed individuals provide the writing data used to construct a library of person-specific adaptation modules, $\mathcal{L} = \{\Delta_1, \ldots, \Delta_K\}$. Each adaptation Δ_i captures source-specific variation through the response distribution induced by $f_{\theta_0+\Delta_i}$ under a common set of generation instructions. Given this fixed library, our goal is to construct additional agent configurations whose aggregate text distribution P_{synth} approximates the human distribution P_{real}. New agents are personalized solely by reusing the learned modules, without collecting additional source texts or fitting a separate adaptation for each new agent. The resulting population should preserve human-like variation while reducing direct attribution to any individual source.

Formally, a sampling distribution q over combinations of modules in $\mathcal{L}$, with each draw specifying a fixed agent configuration. Repeated generations capture within-agent variation, while independently sampled configurations capture between-agent variation. We seek a population satisfying four properties. **1) distributional fidelity**: its pooled text distribution remains close to human writing. **2) population diversity**: agents induce distinct conditional response distributions, allowing more than K behaviorally distinct agents. **3) agent coherence**: a fixed configuration yields consistent behavior across generations. **4) source obfuscation**: generated text is no more attributable to contributing sources than under a permutation reference. Section 3.5 operationalizes these properties and evaluates which are jointly satisfied.

3.2 MODULE LEARNING

For each source individual, we treat their texts as target responses to a neutral writing instruction and optimize next-token cross-entropy over the response tokens. We update only the LoRA parameters (Hu et al., 2022) while keeping the shared backbone f_{θ_0} fixed. For each adapted projection, the source-specific update is $(\alpha/r)B_iA_i$, with rank r and scaling factor α. Training each source independently yields a module library $\mathcal{L} = \{\Delta_1, \ldots, \Delta_K\}$ whose adapted projections are aligned in layer and projection type across sources. All texts within a domain follow the same preprocessing and training protocol, with dropout and weight decay applied to regularize the adaptations. We additionally track source-text reproduction when comparing training budgets. A single-source agent loads one complete module Δ_i across all adapted layers and keeps it fixed throughout generation. The corresponding single-source population samples agents from the K learned modules. Replicating these modules across additional agents does not increase the number of distinct parameter configurations, so this construction supports at most K distinct agents. We use it as the complete-module reference for population fidelity, response matching, and source attribution.

3.3 CROSS-LAYER COMPOSITION

We partition the adapted layers into B contiguous depth blocks. A recipe $\mathbf{s} = (s_1, \ldots, s_B)$ assigns one source module to each block, yielding

$$\Delta_{\mathbf{s}} = \sum_{b=1}^{B} M_b \Delta_{s_b},$$

where M_b selects the adapted projections in block b. For each recipe, s_b is sampled independently and uniformly from $1, \ldots, K$ with replacement. The selected LoRA factors and their scaling are used unchanged within each block, and the resulting recipe remains fixed for the agent throughout generation. Figure 2 illustrates this construction.

This sampling law preserves the marginal distribution over source modules within each block while removing the original dependence of source identity across blocks. It induces K^B possible recipes. When $B = 1$, the construction reduces to sampling complete source modules; for $B > 1$, blocks from different sources can be recombined, expanding the space of agent configurations beyond the K

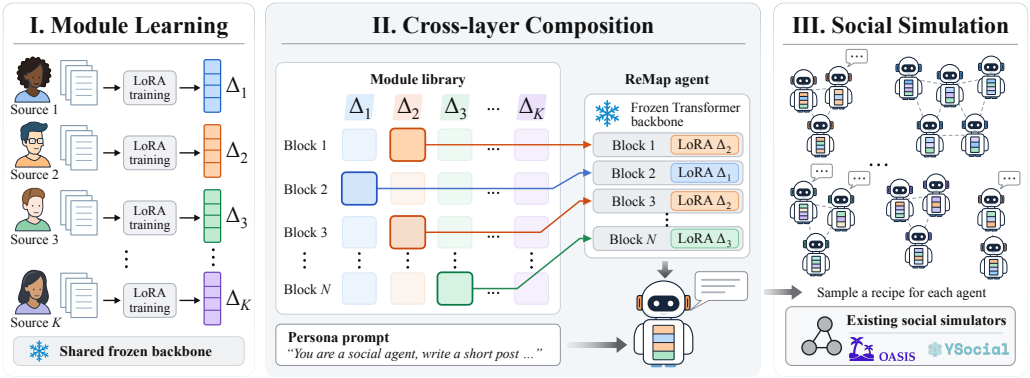

Figure 2. The overview of ReMap. Person-specific LoRA modules are learned from individuals' writing (I); agents compose them across depth blocks into fixed agent recipes (II), forming the simulation population (III).

fitted modules. We determine whether these parameter-level combinations yield distinct behaviors from their generated responses.

Cross-layer composition preserves source-derived variation while reducing how closely a synthesized agent matches any single source module. Because the induced response distribution depends on interactions among the selected blocks, we evaluate how block granularity and cross-block dependence affect population fidelity, agent diversity, coherence, and source attribution. Prior work (Thillainathan et al., 2026) has used layer-wise adapter mixing to reproduce a specified author's style; in contrast, our sampling law generates a population of fixed recipes without fitting a separate target adaptation for each new agent.

A materialized recipe contains the same adapted projections and per-projection rank as a single source module and can therefore be served through a standard adapter interface. Its runtime and storage costs depend on how recipes are materialized and managed by the serving system.

3.4 Deployment in Social Simulation

LLM social simulators such as OASIS and YSocial run a population on a shared platform in discrete rounds. In each round, the environment selects active agents, builds a feed for each from its recommendation system, and asks the language model for an action such as posting, commenting or liking, which then updates the shared timeline. Existing simulators give every agent the same backbone and place its identity in a persona prompt. ReMap agents carry their identity in the backbone weights. Each recipe is materialized as a standard low-rank adapter whose blocks copy the selected source modules exactly, and a multi-adapter server such as vLLM routes every request of an agent to its own adapter. Every population runs under the same prompts, feeds, and action space, and the simulator only maps each agent identifier to an adapter name. For large populations, our engine composes recipes on the fly from the shared library without materializing them (Section 4.6, Algorithm 1).

3.5 Evaluation

We evaluate our approach across fidelity, diversity, coherence, and source obfuscation. To measure **distributional fidelity**, we use the classifier two-sample test (C2ST) (Lopez-Paz & Oquab, 2017) and MAUVE. Both are calibrated against matched disjoint human samples, where lower C2ST and higher MAUVE indicate closer agreement with the human distribution. For **population diversity**, we assess aggregate variation using the Vendi score and the cosine gap to human text (ΔC). We also employ a response matching task, a closed-set problem analogous to speaker identification and authorship attribution (Campbell, 1997; Stamatatos, 2009), to verify that fixed agent configurations induce reproducible conditional responses. **Agent coherence** is evaluated through sentence-level consistency within individual posts and recipe agreement across independent generations. To quantify **source obfuscation**, an author classifier trained on real texts measures how often generated responses are attributed to their source modules compared to a random permutation reference. Finally, we test evaluator sensitivity using length C2ST and an alternative discriminator, and we ex-

Table 1. Population metrics. **V** denotes Vendi and **M** denotes MAUVE.

Model	Condition	Reddit				SCOTUS				Congress			
		C2ST↓	V↑	ΔC↓	M↑	C2ST↓	V↑	ΔC↓	M↑	C2ST↓	V↑	ΔC↓	M↑
Qwen3-4B	Human	52.0	14.0	0.0	95.7	52.0	15.1	0.0	93.8	51.0	18.0	0.0	92.0
	Base	100.0	6.2	13.0	0.6	100.0	4.4	20.5	0.6	100.0	6.7	16.9	0.6
	Generic	99.7	10.6	2.3	1.1	100.0	10.8	**2.9**	0.6	99.4	9.8	9.6	0.8
	Matched	99.6	8.8	7.6	3.1	99.7	9.1	7.7	1.6	98.9	9.0	11.8	1.6
	ReMap	**81.4**	**13.2**	**0.8**	**77.8**	**76.9**	**11.6**	5.0	**71.8**	**82.0**	**13.6**	**5.1**	**41.5**
Gemma-3-4B	Human	52.0	14.0	0.0	95.5	52.0	15.1	0.0	93.8	51.0	18.0	0.0	92.0
	Base	100.0	4.3	18.5	0.5	100.0	8.1	7.7	0.7	100.0	4.3	24.3	0.6
	Generic	99.8	9.5	4.8	0.8	100.0	**13.6**	**−1.1**	0.6	99.7	7.8	13.6	0.8
	Matched	99.8	8.5	8.6	2.8	99.9	11.0	4.4	1.1	99.1	6.5	18.4	2.1
	ReMap	**77.9**	**13.2**	**1.1**	**86.0**	**80.6**	11.2	5.8	**72.7**	**77.3**	**13.7**	**5.2**	**56.2**
Qwen3.5-27B	Human	52.0	14.0	0.0	94.8	52.0	15.1	0.0	96.0	51.0	18.0	0.0	90.7
	Base	99.9	7.1	10.4	0.9	100.0	6.0	15.3	0.6	99.9	5.0	23.8	0.6
	Generic	99.6	9.6	4.3	0.7	99.9	9.0	7.3	0.7	99.6	10.3	9.2	0.6
	Matched	99.7	7.5	10.6	1.8	99.8	7.4	12.2	2.7	99.3	8.3	13.9	1.5
	ReMap	**77.7**	**13.0**	**1.3**	**80.3**	**75.5**	**11.1**	**5.9**	**78.1**	**77.7**	**13.6**	**5.1**	**48.9**
Gemma-4-31B	Human	52.0	13.9	0.0	95.2	52.0	15.1	0.0	93.8	51.0	18.0	0.0	92.0
	Base	100.0	5.7	12.5	0.5	100.0	4.3	20.5	0.5	100.0	4.9	22.2	0.5
	Generic	99.6	9.6	3.8	0.6	99.8	9.3	5.9	0.8	99.7	9.5	9.6	0.6
	Matched	99.7	7.9	9.2	1.9	99.9	7.4	11.9	1.1	99.1	8.0	13.9	1.2
	ReMap	**73.6**	**13.9**	**−0.1**	**86.3**	**77.3**	**11.5**	**5.2**	**65.4**	**76.9**	**13.9**	**4.7**	**52.8**
Qwen3.5-9B	Human	52.0	14.0	0.0	95.8	52.0	15.1	0.0	96.6	51.0	18.0	0.0	92.0
	Base	99.8	8.2	9.0	1.3	99.7	7.8	11.2	1.7	99.8	5.5	21.6	1.0
	Generic	99.5	10.5	2.9	0.8	99.8	**11.8**	**2.2**	1.3	99.1	11.0	7.9	0.9
	Matched	99.7	8.8	8.0	3.6	99.6	8.9	8.7	3.1	99.0	7.5	15.8	2.4
	ReMap	**79.7**	**14.2**	**−0.4**	**71.9**	**78.2**	11.0	6.2	**71.2**	**78.6**	**14.1**	**4.7**	**44.2**
Gemma-3-12B	Human	52.0	14.0	0.0	94.1	52.0	15.1	0.0	93.8	51.0	18.0	0.0	92.1
	Base	100.0	5.0	16.1	0.5	100.0	5.8	15.2	0.5	100.0	3.0	31.4	0.6
	Generic	99.5	8.8	6.5	0.8	99.9	11.1	**2.9**	0.7	99.6	7.7	14.0	0.8
	Matched	99.4	8.7	8.0	4.2	99.9	9.5	7.4	1.1	98.8	7.6	15.1	2.8
	ReMap	**77.5**	**12.8**	**1.6**	**79.4**	**76.5**	**11.5**	5.2	**78.3**	**75.2**	**13.9**	**4.9**	**61.1**

amine stance and information responses in downstream interactions. Full protocols and theoretical derivations are detailed in Appendices A.4, C, and D.4.

4 Experiments

We organize the experiments around four research questions. **RQ1**: Do ReMap populations better match human writing distributions? **RQ2**: Can ReMap create distinguishable agents beyond the source-library size? **RQ3**: Why do ReMap populations match human writing more closely? **RQ4**: How does population construction affect collective behavior in social simulation?

4.1 Settings

We evaluate our approach using corpora from Reddit ChangeMyView, Supreme Court oral arguments (Chang et al., 2020), and congressional posts (Litel, 2017). The primary experiments compare ReMap against unadapted models, generic occupation personas, and distilled writing personas across multiple backbones from the Qwen3 and Gemma families ranging from 4B to 31B parameters (Yang et al., 2025; Team et al., 2026). A secondary comparison adds verbalized sampling (Zhang et al., 2025) to the persona baseline. Model training relies on Low-Rank Adaptation. Details regarding data, parameters, and metrics are available in the Appendix A.

Table 2. Context comparison on Qwen3-4B. Rows change the source material in context, and the last row trains it into a module. Length is C2ST on word counts.

Condition	Reddit			SCOTUS			Congress		
	C2ST↓	Length↓	V↑	C2ST↓	Length↓	V↑	C2ST↓	Length↓	V↑
Human reference	49.7	49.7	17.7	49.8	48.2	18.0	47.7	52.3	21.9
Base	100.0	53.3	6.8	100.0	59.5	4.7	100.0	60.5	7.6
Matched persona	99.9	57.2	9.6	99.9	59.9	10.2	99.4	56.8	9.9
Context, 16 posts	92.7	55.5	12.4	96.3	62.2	11.6	97.6	62.0	10.5
Representative context, 16 posts	93.9	56.8	9.6	96.3	62.7	10.6	98.2	63.8	7.6
Full history, 160 posts	93.1	58.1	12.4	97.0	62.9	13.1	98.2	60.8	9.0
Persona and full history	95.2	58.2	12.0	98.0	62.4	13.5	98.2	59.0	9.4
Single-source LoRA	79.3	56.9	14.4	69.0	50.9	15.3	77.9	57.2	17.4

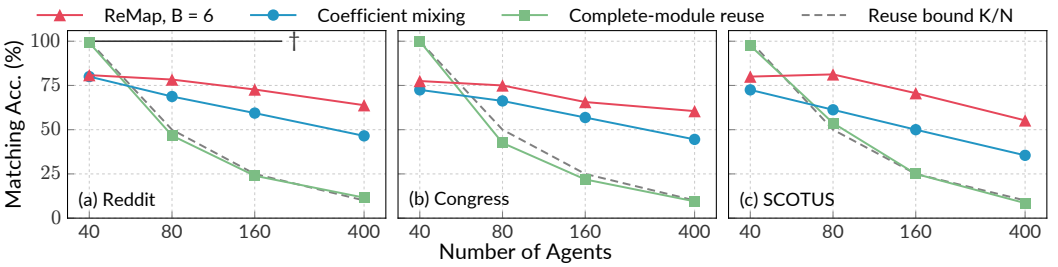

Figure 3. Response matching as populations grow from forty source modules on Qwen3-4B. Reddit averages three seeds, and † marks 200 independent people.

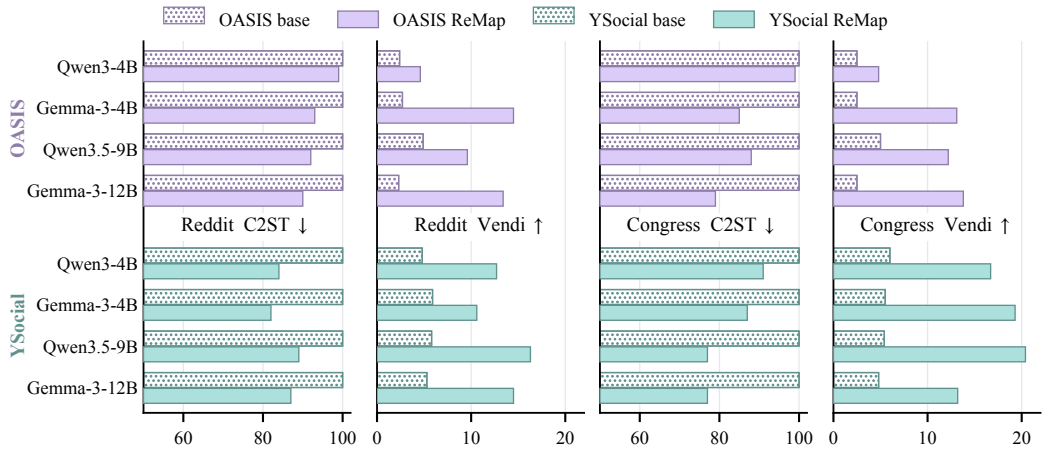

Figure 4. Population measurements inside OASIS and YSocial. Dotted white bars are base agents and filled bars are ReMap agents.

4.2 Population Fidelity (RQ1)

Table 1 compares ReMap populations with the base model and the two persona baselines. ReMap populations achieve lower C2ST and higher MAUVE across all 18 domain–backbone combinations, indicating closer agreement with human writing. On Reddit with Qwen3-4B, for example, C2ST falls from 99.6 to 81.4, and MAUVE rises from 3.1 to 77.8 relative to matched personas. Higher Vendi scores and smaller absolute cosine gaps relative to the base further indicate reduced homogeneity in the generated texts. Persona prompting also increases diversity, but its high C2ST and low MAUVE show that this variation remains substantially different from human writing. Our method improves both distributional agreement and population-level variation.

The gains extend across Reddit discussions, Supreme Court oral arguments, and congressional posts, demonstrating effectiveness across the three domains. The same pattern holds across backbones ranging from 4B to 31B parameters. Even the 27B and 31B base and persona populations remain readily distinguishable from human writing, while ReMap populations retain clear improvements.

These results show that our method brings generated populations closer to human writing across domains and backbone scales while recovering more of the variation present in human text.

4.3 Population Expansion (RQ2)

Although ReMap populations approach human post distributions, a population built from a limited set of source people may still contain far fewer distinct individuals than agents. To resolve this, we grow populations from 40 to 400 agents on the same forty modules in each domain, have every agent answer the same questions in two independent sessions, and test whether response matching can pair each agent's two sessions. As a baseline, complete-module reuse gives each agent one of the forty modules, so copies of a module share one response distribution, which bounds any matcher at K/N. As shown in Figure 3, complete-module reuse follows this bound, while ReMap stays far above it once the population exceeds the library and still reaches 55 to 64 percent matching accuracy at 400 agents in all three domains. Coefficient mixing of three complete modules also exceeds the bound, and ReMap's lead over it widens as the population grows. ReMap populations therefore consist of distinct and stable individuals at ten times the size of their source library. A small library of observed individuals thus becomes a generative basis for societies an order of magnitude larger, decoupling the scale of social simulation from the number of people whose posts must be collected. Appendix C details the protocol and further measurements.

4.4 Drivers of Fidelity Gains (RQ3)

To further investigate why ReMap populations are closer to human writing, we separate three ingredients that every module carries, namely the human text it learns from, the distinction between people, and the storage of that text in the weights. To be more specific, on Qwen3-4B in all three domains, we replace each trained update with a random update of the same norm, load one shared module into every agent, and give the model the same source posts in context without training on them. As shown in Figure 5, random updates leave C2ST at the base level of 100 in every domain, whereas any module trained on human text lowers it. A shared module already closes part of the gap, and the one fitted to all source text closes most of it on Reddit and SCOTUS, yet the lowest C2ST in every domain comes from a library of distinct modules. Placing the same posts in context tests the last ingredient, and Table 2 shows that even the full 160-post history keeps C2ST at 93.1 or above, while the module trained on these posts reaches 69.0 to 79.3. These findings demonstrate that the realism of a simulated population comes from person-specific parameters learned from human writing, the very components that ReMap recombines into new individuals.

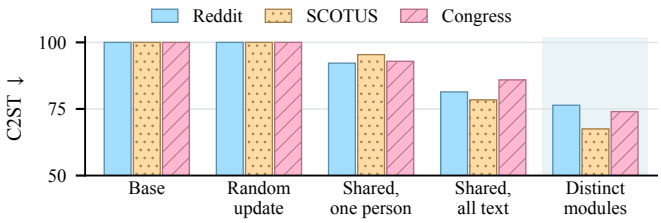

Figure 5. C2ST on Qwen3-4B as ingredients of the learned modules are removed, one bar per domain.

4.5 Collective Behavior (RQ4)

To determine whether our method generalizes to simulated social media scenarios and reproduces real social phenomena, we conduct downstream simulation experiments. We integrate ReMap agents into two social media simulators, YSocial and OASIS, and observe that ReMap benefits downstream tasks in both environments. First, we examine whether agents maintain human-like writing patterns when processing feeds, responding to peers, and utilizing platform tools. We evaluate both platforms using four model backbones across Reddit and Congress populations while preserving each simulator's original interaction mechanics. As Figure 4 illustrates, a classifier perfectly distinguishes base agent outputs from human posts with a C2ST of 100 in every configuration. Additionally, the low Vendi score for base agents indicates that the simulated population exhibits limited generative diversity. In contrast, ReMap agents lower C2ST and increase the Vendi score across all sixteen configurations. Ultimately, these results demonstrate that ReMap robustly handles the dynamic complexity of full social simulations and generates highly authentic multi-agent populations.

Second, we examine whether the opinions agents hold after interaction match those of real people. We seed OASIS with six contested claims from ChangeMyView, keep each claim at the top of

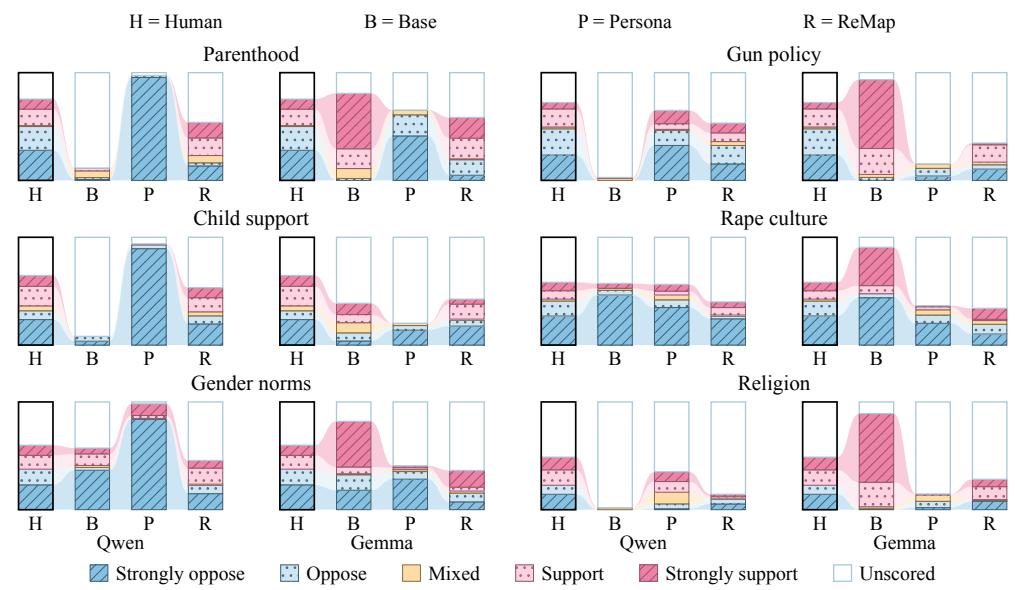

Figure 6. Human and simulated stance distributions after twelve OASIS rounds on six ChangeMyView claims and two backbones. Ribbons connect matching stance categories.

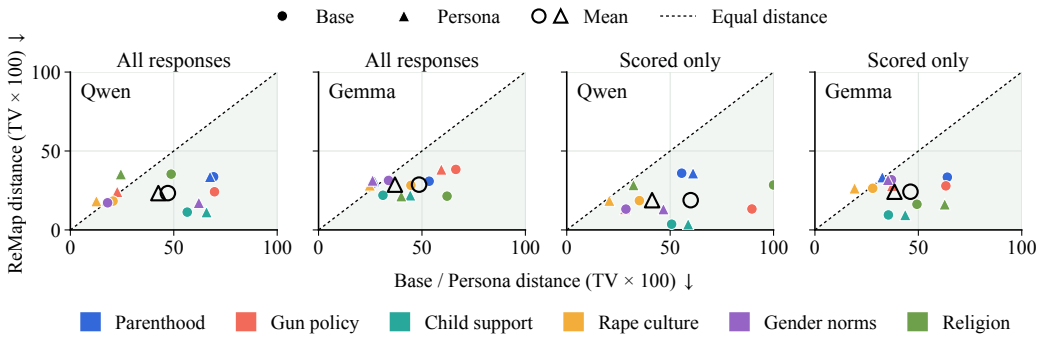

Figure 7. Distance to the human stance distribution in OASIS on two backbones. Points below the diagonal favor ReMap, and hollow symbols give six-topic means.

every feed as the opening post of a thread, and let 111 agents discuss it for twelve rounds with conversation memory kept. A judge model then labels the stance of each final-round post and of the human comments on the same claim, and we compare the two stance distributions on Qwen3-4B and Gemma-3-4B. As shown in Figure 6, base agents either withhold a stance or crowd onto a single side, and persona prompts shift the population without recovering the human mix of positions. Figure 7 places ReMap agents closest to humans on both backbones, at average distances of 0.23 and 0.29, and the advantage holds when only posts that take a stance are compared. Appendix E.2 details the protocol. These results indicate that ReMap populations reproduce how real people divide on contested issues, the very quantity that opinion simulations are built to measure.

Third, we examine the collective dynamics that social simulation studies report, namely how agents respond to the majority and to false information. In OASIS with conversation memory kept, agents discuss a contested vaccine mandate for twelve rounds, and we count a change of stance as conformity when it moves toward the preceding majority. In a separate study, agents discuss three false health claims, and we record how often their posts endorse each claim. As shown in Table 3, base agents follow the majority in nearly every change of stance and end with almost identical views. Persona prompts weaken this pull only partly, whereas ReMap agents conform least and keep the widest spread of stances. Responses to false claims follow the same contrast, since base and persona agents almost never endorse them, while ReMap populations contain an endorsing subgroup of 11 to 18 percent, close to the shares that public surveys report for the same beliefs. These findings show that population construction determines the conformity and susceptibility a simulation reports,

Table 3. Collective responses in OASIS. Conformity of 0.5 means no majority pull, and human rates are survey shares from Gallup (2020) and Carleton University (2020).

Population	Opinion interaction			Population	Information response		
	Entropy↑	Conformity↓	Eligible		MMR autism	5G COVID	Vitamin C COVID
Base	0.307	0.991	569	Base	0.00	0.00	0.00
Persona	0.676	0.803	664	Persona	0.01	0.00	0.00
Single-source	1.249	0.763	615	Human survey	0.10	0.11	–
ReMap	1.278	0.684	722	**ReMap**	0.18	0.11	0.12

Table 4. Serving measurements on Qwen3-4B. Agents are generated in batches of 16.

Sources	Source library			Agents	Agent population		
	Dense (×)↓	Gather (×)↓	GPU (GB)↓		Time (s)↓	ms/agent↓	GPU (GB)↓
40	1.44	1.22	11.2	1,000	471	471	13.0
200	3.60	1.22	21.9	3,000	1,293	431	13.0
1,000	15.34	1.22	75.8	10,000	4,314	431	13.0

and ReMap agents restore the dissenting minority and the susceptible subgroup that uniform agent societies cannot express.

4.6 Serving Cost

One possible concern is that ReMap may demand substantial resources, since a population can hold thousands of distinct recipes. Thanks to the development of LLM serving systems such as vLLM and S-LoRA (Kwon et al., 2023; Sheng et al., 2023), a single frozen backbone can serve many low-rank adapters at once. Building on this design, we implement an engine that composes each agent's modules layer by layer in one forward pass. Where a naive implementation computes every module in the library for every token, our engine gathers only the modules each agent activates, keeping outputs identical and compute independent of library size. Since each agent is only a routing table over the shared library, memory depends on the number of sources and stays independent of the number of agents. As shown in Table 4, the dense compute multiplier grows with every added source, while gathering keeps it constant. Memory grows gradually with the library, yet a few hundred sources are ample, since all our experiments run on forty sources per domain within 13.0 GB. As the population grows from 1,000 to 10,000 agents, generation time rises linearly while memory stays at 13.0 GB. ReMap therefore decouples population size from hardware cost and brings large-scale social simulation within reach of a single GPU. Appendix F gives the gathering algorithm, hardware settings, host-memory paging, and the cost of persona and context conditioning.

5 Conclusion

In this paper, we proposed ReMap to construct diverse LLM populations for social simulation from observed individual variation. By learning person-specific adaptation modules from real-world data and recombining them across model depth into fixed configurations, our method enables a finite source library to generate a much larger agent population. We evaluated this framework across three domains and six LLM backbones, together with downstream simulations in OASIS and YSocial. Our results demonstrate that ReMap populations better match human distributions while preserving substantially greater variation than persona-prompted agents. Moreover, expanding a small set of source modules into hundreds of behaviorally distinguishable agents yields collective responses that better capture disagreement, minority opinions, and heterogeneous information responses. These findings suggest that population heterogeneity can be recovered by learning and recombining individual variation. Ultimately, our work establishes cross-layer composition as a scalable foundation for constructing more representative LLM populations for social simulation.

AI USE STATEMENT

AI assistants supported code development and language editing during this project. Language models also generate the simulated populations and label text in the behavioral evaluations. Those experimental uses follow the procedures described in the method and evaluation protocols.

ETHICS STATEMENT

The study uses public research corpora and public-record text. Contributor-level adaptation and synthetic social responses introduce risks of impersonation and coordinated influence. The intended release comprises training and evaluation code with aggregate results. Person-level data and source adapters are excluded from that intended release, and ReMap-agent attribution is evaluated as an empirical property. Removing a contributor requires removing their source module and rebuilding materialized recipes that include it.

REPRODUCIBILITY STATEMENT

Our code and data are anonymously available at `https://bit.ly/4iHRpBb`, with one runner per experiment that uses the settings reported in the paper. The release covers data construction from the public corpora, module training, ReMap composition, the OASIS and YSocial integrations and the evaluation metrics. Appendix A lists the data, prompts, training and generation settings, and Appendix A.4 specifies the estimators.

REFERENCES

Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of ICML*, pp. 337–371. PMLR, 2023.

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023.

Luke Baines, Anton Gonzalvez Hawthorne, Mariia Koroliuk, Irakli Shalibashvili, Clément Dumas, Konstantinos Voudouris, and David Demitri Africa. Persona cartography: Charting language model personality traits in weight space. *arXiv preprint arXiv:2607.07916*, 2026.

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

Joseph P. Campbell. Speaker recognition: A tutorial. *Proceedings of the IEEE*, 85(9):1437–1462, 1997.

Yong Cao, Haijiang Liu, Arnav Arora, et al. Specializing large language models to simulate survey response distributions for global populations. In *Proceedings of NAACL*, pp. 3141–3154, 2025.

Carleton University. New Carleton study finds COVID-19 conspiracies and misinformation spreading online, 2020. Survey of 2,000 Canadian residents with Abacus Data, May 5–8, 2020.

Jonathan P. Chang, Caleb Chiam, Liye Fu, et al. ConvoKit: A toolkit for the analysis of conversations. In *Proceedings of SIGDIAL*, pp. 57–60, 2020.

Nuo Chen, Yicheng Tong, Yuzhe Yang, et al. Diversity collapse in multi-agent LLM systems: Structural coupling and collective failure in open-ended idea generation. In *Findings of ACL*, 2026.

Roger Creus Castanyer, Geoffrey Bradway, Lorenz Wolf, Maxwill Lin, Augustine N. Mavor-Parker, and Matthew James Sargent. Populora: Co-evolving llm populations for reasoning self-play. *arXiv preprint arXiv:2605.16727*, 2026.

Ricardo Dominguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dünner. Questioning the survey responses of large language models. In *Proceedings of NeurIPS*, 2024.

Gallup. Fewer in U.S. continue to see vaccines as important. `https://news.gallup.com/poll/276929/fewer-continue-vaccines-important.aspx`, 2020. Survey of 1,025 U.S. adults, December 2–15, 2019.

Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huandong Wang, Depeng Jin, and Yong Li. S3: Social-network simulation system with large language model-empowered agents. *arXiv preprint arXiv:2307.14984*, 2023.

Minghao Guo, Ziyi Ye, Wujiang Xu, et al. Individual turing test: A case study of LLM-based simulation using longitudinal personal data. In *Proceedings of SIGIR*, pp. 3770–3775, 2026.

Edward J. Hu, Yelong Shen, Phillip Wallis, et al. LoRA: Low-rank adaptation of large language models. In *Proceedings of ICLR*, 2022.

Tiancheng Hu and Nigel Collier. Quantifying the persona effect in llm simulations. In *Proceedings of ACL*, pp. 10289–10307, 2024.

Zhengyu Hu, Jianxun Lian, Zheyuan Xiao, et al. Population-aligned persona generation for LLM-based social simulation. *arXiv preprint arXiv:2509.10127*, 2025.

Wenyue Hua, Lizhou Fan, Lingyao Li, Kai Mei, Jianchao Ji, Yingqiang Ge, Libby Hemphill, and Yongfeng Zhang. War and peace (waragent): Large language model-based multi-agent simulation of world wars. *arXiv preprint arXiv:2311.17227*, 2023.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, et al. Editing models with task arithmetic. In *Proceedings of ICLR*, 2023.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of SOSP*, pp. 611–626, 2023.

Jack Lanchantin, Angelica Chen, Shehzaad Dhuliawala, et al. Diverse preference optimization. *arXiv preprint arXiv:2501.18101*, 2025.

Alex Litel. congresstweets: Datasets of the daily Twitter output of Congress. `https://github.com/alexlitel/congresstweets`, 2017. MIT License. Daily archives from 2017-06-21; last daily update 2023-07-11.

Xinyue Liu, Harshita Diddee, and Daphne Ippolito. Customizing large language model generation style using parameter-efficient finetuning. In *Proceedings of INLG*, pp. 412–426, 2024.

David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. In *Proceedings of ICLR*, 2017.

Marlene Lutz, Indira Sen, Georg Ahnert, Elisa Rogers, and Markus Strohmaier. The prompt makes the person (a): A systematic evaluation of sociodemographic persona prompting for large language models. In *Findings of EMNLP*, pp. 23212–23237, 2025.

Joon Sung Park, Joseph O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of UIST*, 2023.

Joon Sung Park, Carolyn Q. Zou, Jonne Kamphorst, et al. LLM agents grounded in self-reports enable general-purpose simulation of individuals. *arXiv preprint arXiv:2411.10109*, 2024.

Dipkumar Patel. Representational collapse in multi-agent LLM committees: Measurement and diversity-aware consensus. *arXiv preprint arXiv:2604.03809*, 2026.

Jinghua Piao, Yuwei Yan, Jun Zhang, et al. AgentSociety: Large-scale simulation of LLM-driven generative agents advances understanding of human behaviors and society. *arXiv preprint arXiv:2502.08691*, 2025.

Akshara Prabhakar, Yuanzhi Li, Karthik Narasimhan, et al. LoRA soups: Merging LoRAs for practical skill composition tasks. In *Proceedings of COLING: Industry Track*, pp. 644–655, 2025.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *Proceedings of ICML*, pp. 29971–30004, 2023.

Yunfan Shao, Linyang Li, Junqi Dai, et al. Character-LLM: A trainable agent for role-playing. In *Proceedings of EMNLP*, pp. 13153–13187, 2023.

Ying Sheng, Shiyi Cao, Dacheng Li, et al. S-LoRA: Serving thousands of concurrent LoRA adapters. *arXiv preprint arXiv:2311.03285*, 2023.

Efstathios Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556, 2009.

Joseph Suh, Erfan Jahanparast, Suhong Moon, et al. Language model fine-tuning on scaled survey data for predicting distributions of public opinions. In *Proceedings of ACL*, pp. 21147–21170, 2025.

Seungjong Sun, Eungu Lee, Dongyan Nan, et al. Random silicon sampling: Simulating human sub-population opinion using a large language model based on group-level demographic information. *arXiv preprint arXiv:2402.18144*, 2024.

Zhaoxuan Tan, Zheyuan Liu, and Meng Jiang. Personalized pieces: Efficient personalized large language models through collaborative efforts. In *Proceedings of EMNLP*, pp. 6459–6475, 2024a.

Zhaoxuan Tan, Qingkai Zeng, Yijun Tian, et al. Democratizing large language models via personalized parameter-efficient fine-tuning. In *Proceedings of EMNLP*, pp. 6476–6491, 2024b.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Himanshu Thakur, Eshani Agrawal, and Smruthi Mukund. Personas within parameters: Fine-tuning small language models with low-rank adapters to mimic user behaviors. *arXiv preprint arXiv:2509.09689*, 2025.

Sarubi Thillainathan, Ji-Ung Lee, Michael Sullivan, et al. AuthorMix: Modular authorship style transfer via layer-wise adapter mixing. *arXiv preprint arXiv:2603.23069*, 2026.

Hexi Wang, Yujia Zhou, Bangde Du, Qingyao Ai, and Yiqun Liu. Parametric social identity injection and diversification in public opinion simulation. In *Proceedings of KDD*, pp. 4894–4905, 2026.

Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, et al. Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *Proceedings of ICML*, pp. 23965–23998, 2022.

Zengqing Wu, Run Peng, Takayuki Ito, Makoto Onizuka, and Chuan Xiao. LLM-based social simulations require a boundary. *arXiv preprint arXiv:2506.19806*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Ziyi Yang, Zaibin Zhang, Zirui Zheng, et al. OASIS: Open agent social interaction simulations with one million agents. *arXiv preprint arXiv:2411.11581*, 2024.

Jinyi Ye, Lei Cao, Ding Chen, and Emilio Ferrara. Stop drawing scientific claims from LLM social simulations without robustness audits. *arXiv preprint arXiv:2605.18890*, 2026.

Jiayi Zhang, Simon Yu, Derek Chong, et al. Verbalized sampling: How to mitigate mode collapse and unlock LLM diversity. *arXiv preprint arXiv:2510.01171*, 2025.

Appendix

Content

A Experimental Settings **13**
A.1 Data 13
A.2 Conditions and Prompts 14
A.3 Training and Generation 15
A.4 Metrics 16
B Population Fidelity **17**
B.1 Other Evaluators 17
B.2 Sampling Variation 17
B.3 Verbalized Sampling 17
C Population Expansion **18**
C.1 Protocol 18
C.2 Matching Bound 19
C.3 Results at 400 Agents 19
C.4 Matcher Range 19
D ReMap Analysis **20**
D.1 Context Conditioning 20
D.2 Module Ingredients 20
D.3 Block Count 21
D.4 Agent Coherence 21
D.5 Text Reproduction 22
E Simulation Studies **22**
E.1 Platform Populations 22
E.2 OASIS Stance 23
E.3 Opinion Interaction 23
E.4 Information Responses 24
F Serving Cost **24**

A Experimental Settings

A.1 Data

The three corpora are Reddit ChangeMyView comments and Supreme Court oral arguments from ConvoKit (Chang et al., 2020) alongside posts by members of Congress from the congresstweets archive (Litel, 2017). Each domain supplies 40 source individuals, and Table 5 summarizes their training text and the human evaluation pool. Reddit sources consist of the 40 users with the most comments ranging from 60 to 1,200 characters, selected after removing deleted accounts, moderator bots and comments that open with a quotation. Each user contributes their first 160 such comments, truncated to 600 characters. SCOTUS sources comprise 40 advocates with up to 160 utterances of 80 to 1,200 characters each. Congress sources include the 40 member accounts with the highest volume of posts after removing retweets, links, quoted posts and leading mentions, restricting the selection to posts of 8 to 60 words. Each account contributes 160 randomly drawn posts. Congressional training posts originate from an earlier archive download, while the congressional evaluation pool spans the window 13 January to 11 July 2023.

Every evaluation pool is written by individuals outside the 40 sources and shares no overlap with the training posts. The Reddit pool contains 2,000 comments from 499 other users, the SCOTUS pool encompasses 1,000 utterances from 634 other speakers, and the Congress pool includes 50 posts from each of the next 40 members ranked by volume. Human and generated texts undergo identical cleaning, which removes quoted lines and moderator templates while retaining texts of 40 to 1,000 characters.

Table 5. Training text and human evaluation pools. Word counts are means per text.

	Reddit	SCOTUS	Congress
Source individuals	40	40	40
Training posts	6,400	6,026	6,400
Posts per source	160	122 to 160	160
Words per training post	61.7	63.0	30.8
Evaluation texts after cleaning	2,000	960	1,999
Evaluation authors	499	634	40
Words per evaluation text	67.5	55.9	32.6

A.2 Conditions and Prompts

All conditions share the backbone and the generation request, and they differ in what each agent is given. The base condition runs the unadapted backbone with the system instruction `You are a social-media user.` The generic persona condition cycles through sixteen occupation descriptions. The matched persona condition provides each agent with a short second-person description of one source individual's views, stances, tone and writing style, distilled from their posts. Qwen3-4B writes the Reddit and SCOTUS descriptions using the instruction shown below. The single-source condition loads one complete source module, and the ReMap condition loads a cross-layer recipe. Single-source and ReMap agents keep the base system instruction, differing from the base strictly in their weights.

Two further constructions appear in the supplementary comparisons. Coefficient mixing loads three complete modules at once,

$$\Delta W = \sum_{j=1}^{K} c_j \frac{\alpha}{r} B_j A_j,$$

where the three selected modules share one coefficient, $1/3$ or $1/\sqrt{3}$, and all others remain zero. Verbalized sampling (Zhang et al., 2025) asks a matched-persona agent for several candidate posts with probabilities and samples one of them, utilizing the published instruction and its tail-sampling variant.

The prompt cards below give the literal strings, with variables in braces.

Agent instructions

System instruction for base, single-source and ReMap agents

```
You are a social-media user.
```

Generic persona, two of sixteen

```
You are a 52-year-old union electrician, blunt, working-class,
skeptical of elites.
You are a 24-year-old grad student in philosophy, precise, fond of
thought experiments.
```

Generation request, with the target word count sampled from the human pool

```
{description} Aim for about {n} words.  Output only the text.
```

Description on Reddit, SCOTUS and Congress

```
Write a Reddit comment in your own voice -- an opinion, argument, or
experience you would actually post.
Make a brief remark in your own voice, as in a Supreme Court oral
argument.
Write a short tweet in your own voice -- a political opinion,
statement, or reaction you would actually post.
```

Matched persona distillation

System

```
You are a careful analyst who reads a person's real posts and writes a
short persona description of them.
```

User, with eighteen of the person's posts cut to 400 characters each

```
Below are real posts by ONE person, separated by ---.  {posts} Write a
2-3 sentence persona describing this person's main views, typical
stances, tone, and writing style, in the second person, usable as a
system prompt for role-play.  Begin with "You are someone who".  Output
ONLY the description, no preamble.
```

Expansion questions

Instruction on Reddit, followed by one situation

```
Write a Reddit comment responding to this situation.  Give your own
view and explain your reasoning.  Output only the comment.
```

The SCOTUS and Congress instructions ask for a brief remark in an oral argument and a short tweet from the agent's own account. The twelve situations are the same in every domain. They concern returning to the office, moderating rude comments, public transport against wider roads, trading browsing data for recommendations, apartments in detached-house areas and a school-day phone ban. The rest concern helping few people a lot or many a little, replacing tips with higher prices, duties toward workers facing automation, charging per bag of rubbish, a friend who shares disagreeable opinions, and two reports that reach opposite conclusions. One of them in full

```
A city has money for either more public transport or wider roads, but
not both.  Which investment would you prioritize, and why?
```

OASIS seed posts

Persona of base and ReMap agents

```
You are a user on a social media platform.  Share your honest opinion
in your own voice.
```

Vaccine mandate

```
Vaccines should be mandatory for everyone, with no personal or
religious exemptions.  What is your honest opinion -- do you agree or
disagree, and why?
```

ChangeMyView claims, each followed by the same question, for example

```
Men should either be able to waive their financial obligations to their
child, or they should have abortion rights.
```

False claim, posted as written, for example

```
Notice how the COVID hotspots line up exactly with the new 5G rollout
maps?  The radiation weakens your immune system and spreads the virus.
They won't tell you this on the news.
```

Stance judges

Vaccine mandate, Claude Haiku 4.5, first 500 characters of each post

```
A user is reacting to the claim 'vaccines should be mandatory for
everyone with no exemptions'.  Label their stance as exactly one word:
pro (supports the mandate / agrees), anti (opposes the mandate /
disagrees / favors personal choice), or neutral (no clear stance or
off-topic).  Read implicit/argumentative stance.  Output one word.
```

ChangeMyView claims, DeepSeek-V4-Flash, with system instruction `You are a precise annotator.  Output only one of:  X -2 -1 0 1 2`

```
Claim:  "{claim}" Comment:  "{text}" First:  does the comment take ANY
position on THIS claim?  If it does NOT (off-topic, only a question, or
about a side-point), answer X. Otherwise rate its stance on the claim:
-2 strongly disagree, -1 disagree, 0 mixed/neutral, 1 agree, 2 strongly
agree.  Answer with ONLY one of:  X -2 -1 0 1 2 Answer:
```

A.3 Training and Generation

Training updates only the LoRA factors on the query, value, output and MLP down projections. Table 6 details the settings. Qwen3-4B modules train for 60 steps and Gemma-3-4B modules for 30, and the larger backbones utilize step counts tailored per backbone and domain so that generated text rarely copies source posts (Appendix D.5).

ReMap splits the ordered decoder layers into contiguous blocks whose sizes differ by at most one layer, defaulting to six blocks.

Table 6. Training and generation settings. Where three values appear, they belong to Reddit, SCOTUS and Congress.

Training	Value	Composition and generation	Value
LoRA rank r	16	Depth blocks	6
Scaling factor α	32	Max new tokens	90
Optimizer	AdamW	Temperature	0.9
Learning rate	3×10^{-4}	Word target, upper bound	180
Batch size	4	Word target, lower bound	15 / 5 / 8
Gradient clipping	1.0	Minimum length factor	1.05 / 0.5 / 1.0
Dropout	0.1	Generation repeats	5
Weight decay	0.01	Texts kept per repeat, at most	150
Posts per source	160		

Target word counts are sampled from the human pool, providing identical targets to every condition. A shared token cap bounds each output, and each condition maintains its own seeded sampling stream. After filtering, we draw equal numbers of human and generated texts within each configuration. Table 7 details the repeats, sample sizes and references for every comparison.

A.4 METRICS

Text metrics utilize the normalized CLS embedding of BGE-large-en-v1.5 with inputs truncated to 128 tokens. Each human reference applies the same procedure to two disjoint human samples.

C2ST balances the human and generated texts, trains a classifier on half of each, and reports AUC on the remaining half. The classifier employs logistic regression on 32 principal components, fitting the projection and normalization on the training half. The AUC is averaged over 20 random splits with tied scores assigned average ranks. Length C2ST executes the same procedure on word counts. MAUVE applies identical embeddings to texts pooled across the generation repeats, incorporating equal numbers of human and generated texts alongside three quantization draws. These pools contain 603 to 624 texts per class on Reddit, 414 or 415 on SCOTUS and 637 to 639 on Congress, utilizing one quantization bucket per ten texts. Fréchet distance compares the means and covariances of the embeddings. Two disjoint human samples of 150 texts yield 0.317 on Reddit, 0.328 on SCOTUS and 0.355 on Congress.

For normalized embeddings E, Vendi is computed from the eigenvalues λ_i of the kernel $EE^\top/n$,

$$\mathrm{Vendi}(E) = \exp\left(-\sum_i \lambda_i \log \lambda_i\right).$$

Mean cosine averages the off-diagonal entries of $EE^\top$, and the cosine gap ΔC in Table 1 represents the generated value minus the human value, multiplied by 100.

Response matching tests whether an agent can be reidentified from its answers, and Appendix C details the procedure and the accuracy bound for reused modules. Sentence coherence averages the cosine similarity between adjacent sentences in posts containing more than one sentence. Recipe consistency matches the same recipes across two independently generated populations (Appendix D.4).

Source attribution relies on an author classifier trained once on real texts of the source individuals. Its features mask content words and retain function words, punctuation and surface statistics. For a recipe drawn from the set of sources S, the rate at which its texts are assigned to a member of S is divided by the chance level $|S|/K$. Because sources are drawn with replacement, $|S|$ can be smaller than the block count. A permutation reference shuffles recipes across generated texts while preserving the classifier outputs. Stance, conformity and endorsement measures are defined alongside their simulation protocols in Appendix E.

Table 7. Generation and measurement protocol of each comparison.

Comparison	Generation and sampling	Measurements
Population, Tables 1 and 11	Five repeats, at most 150 texts per condition and repeat, 90-token cap	C2ST, Vendi, cosine gap and MAUVE against human text
Context, Table 2	Three repeats, source-matched histories, 220-token cap on Reddit and SCOTUS and 70 on Congress	C2ST, length C2ST and Vendi
Expansion, Figure 3	40 to 400 agents, two sessions of 96 answers each; three seeds on Reddit and one on SCOTUS and Congress	Response matching, and C2ST, Vendi and Fréchet distance on 800 free-writing texts per size
Block count, Table 12 and Figure 8	Population protocol for the table; forty recipes with twelve posts each for attribution	Human reference; permutation reference
Platforms, Figure 4	OASIS with 111 agents over twelve rounds and YSocial with 80 ReMap agents over sixteen time slots, one run per configuration	C2ST and Vendi within each configuration
Opinions, Figures 6 and 7 and Table 3	111 or 121 agents, twelve rounds, conversation memory kept, repetition penalty 1.2 except Gemma-3-4B in the stance study	Stance distance, stance entropy, conformity and endorsement
Serving, Tables 4 and 14	Batches of 16 agents, one generated post per agent	Compute multiplier, time and memory

B Population Fidelity

B.1 Other Evaluators

Table 8 replaces the frozen-embedding classifier with two encoders trained end-to-end on the identical generated texts. MPNet trains for three epochs with a batch size of 16, a learning rate of 2×10^{-5} and inputs of up to 256 tokens. ReMap text remains more difficult to distinguish from human text compared to base and persona text under both encoders.

We also put the same question to people. Two annotators rated topic-matched Reddit comments, where each generated comment was paired with a human comment of similar length. Their averaged scores yield an AUC for each population along with bootstrap intervals. Evaluators distinguish base text from human text at an AUC of 98.8 and matched-persona text at 92.6, whereas single-source and ReMap texts drop to 77.7 and 73.9 respectively.

B.2 Sampling Variation

Table 9 reports the standard deviations corresponding to Table 1, calculated across the five generation repeats for C2ST, Vendi and the cosine gap, as well as across quantization and sampling draws for MAUVE. These calculations hold the source library fixed. A separate Reddit run using Qwen3-4B resamples the 40 source individuals 500 times, yielding a C2ST standard deviation of 1.7 points. Leaving out any single individual shifts the estimate by a maximum of 2.8 points. In this same run, unseen posts from the 40 source individuals differ from the general human pool with a C2ST of 54.9.

B.3 Verbalized Sampling

We compare verbalized sampling alongside matched personas across all six backbones, averaged over three seeds. C2ST utilizes the frozen-embedding classifier from Table 1, and the MPNet eval-

Table 8. C2ST with discriminators trained end to end on the same generated texts.

Population	MPNet (↓)	RoBERTa-large (↓)
Human reference	47	49
Base	100	100
Persona	100	100
ReMap	86	88

Table 9. Standard deviations behind Table 1. C2ST, ΔC and M are in percentage points, as in that table.

Model	Condition	Reddit				SCOTUS				Congress			
		C2ST	V	ΔC	M	C2ST	V	ΔC	M	C2ST	V	ΔC	M
Qwen3-4B	Base	0.0	0.11	0.4	0.1	0.0	0.10	0.7	0.0	0.0	0.14	0.4	0.0
	Generic	0.1	0.15	0.4	0.2	0.0	0.39	0.8	0.2	0.2	0.14	0.4	0.1
	Matched	0.2	0.24	0.5	0.4	0.1	0.39	1.0	0.4	0.2	0.24	0.5	0.1
	ReMap	2.9	0.23	0.4	3.2	2.3	0.43	0.8	0.9	2.1	0.34	0.5	2.0
Gemma-3-4B	Base	0.0	0.09	0.5	0.1	0.0	0.48	1.1	0.1	0.0	0.24	1.2	0.1
	Generic	0.1	0.22	0.5	0.1	0.0	0.45	0.7	0.1	0.2	0.27	0.7	0.0
	Matched	0.1	0.35	0.8	0.6	0.0	0.12	0.3	0.2	0.4	0.18	0.7	0.2
	ReMap	4.8	0.23	0.3	1.4	2.9	0.28	0.6	7.9	2.3	0.62	0.8	4.6
Qwen3.5-27B	Base	0.0	0.20	0.5	0.0	0.0	0.19	0.8	0.1	0.1	0.25	1.1	0.0
	Generic	0.1	0.24	0.4	0.1	0.1	0.52	1.2	0.0	0.1	0.24	0.5	0.0
	Matched	0.2	0.18	0.4	0.1	0.1	0.25	0.8	0.4	0.5	0.40	1.0	0.1
	ReMap	4.0	0.29	0.4	2.5	2.1	0.37	0.7	1.6	1.7	0.30	0.4	2.5
Gemma-4-31B	Base	0.0	0.12	0.5	0.1	0.0	0.09	0.4	0.0	0.0	0.18	0.8	0.0
	Generic	0.1	0.26	0.5	0.1	0.1	0.20	0.5	0.1	0.1	0.15	0.4	0.0
	Matched	0.2	0.16	0.4	0.3	0.1	0.26	0.8	0.3	0.2	0.34	0.9	0.1
	ReMap	2.8	0.40	0.7	2.0	2.4	0.47	0.8	4.3	3.1	0.34	0.7	4.1
Qwen3.5-9B	Base	0.0	0.19	0.5	0.2	0.2	0.33	0.9	0.3	0.1	0.22	0.8	0.0
	Generic	0.2	0.16	0.4	0.0	0.1	0.50	1.1	0.2	0.4	0.22	0.4	0.1
	Matched	0.2	0.18	0.5	0.6	0.2	0.22	0.6	0.5	0.4	0.15	0.4	0.4
	ReMap	3.1	0.53	0.7	3.4	4.1	0.40	0.9	2.0	1.7	0.72	1.0	0.8
Gemma-3-12B	Base	0.0	0.22	0.8	0.0	0.0	0.06	0.4	0.0	0.1	0.13	0.9	0.0
	Generic	0.3	0.17	0.4	0.0	0.0	0.33	0.7	0.1	0.1	0.18	0.5	0.1
	Matched	0.1	0.26	0.6	1.1	0.1	0.23	0.5	0.1	0.4	0.30	0.9	0.4
	ReMap	5.2	0.20	0.3	0.9	3.4	0.15	0.3	1.7	1.6	0.17	0.1	4.4

uation employs the trained discriminator from Table 8. Both variants increase Vendi scores in most configurations, whereas C2ST remains at 98.5 or above and the MPNet discriminator stays at 99.2 or above across every measured configuration.

C Population Expansion

C.1 Protocol

The expansion study maintains one library of 40 modules per domain on Qwen3-4B and scales the population to 40, 80, 160 and 400 agents. Complete-module reuse distributes the 40 modules evenly across the agents. ReMap assigns every agent a unique recipe comprising two or six blocks, deliberately excluding recipes that repeat or draw every block from a single individual. On Reddit we replicate the study using three population seeds, 201, 202 and 203. SCOTUS and Congress utilize seed 201 to compare reuse, six-block ReMap and the $1/\sqrt{3}$ mixture.

Each agent answers twelve common questions without an identity prompt, personal history or memory across questions. The agent answers every question eight times in each of two independent sessions, where the first one, two, four or eight answers define the observation budgets. Generation applies a 90-token cap, a temperature of 0.9, a top-p of 0.95 and a batch size of 16, or 128 for SCOTUS and the Congress mixture. The backbone operates in bfloat16, while the adapters run in float32.

For each question, we average an agent's answer embeddings within a session, subtract the gallery mean and concatenate the twelve question vectors into a single profile. Queries are matched to gallery profiles based on cosine similarity, treating ties as uniform guesses. The primary measurement matches the first session to the second at 400 agents with eight answers per question. Style matching repeats this process using writing-style features standardized by gallery statistics. The hu-

man reference for C2ST, Vendi and Fréchet distance is deduplicated, entirely excludes exact overlaps with source texts and is divided into two disjoint samples.

C.2 Matching Bound

For a fixed gallery, let $h_i(x)$ denote the probability that the matcher assigns query x to label i, ensuring these probabilities sum to one for each query. Suppose the N labels fall into M groups G_g whose members share a single answer distribution P_g. Assuming uniformly weighted labels, the expected accuracy satisfies

$$\mu_A = \frac{1}{N} \sum_{g=1}^{M} \int \sum_{i \in G_g} h_i(x)\, dP_g(x) \leq \frac{M}{N}.$$

Reusing K modules under identical inputs guarantees $M \leq K$. Consequently, reuse at 400 agents is theoretically capped at 10 percent in expectation, though a single finite sample might land above this threshold.

To evaluate the pooled Reddit results, we combine the query scores from the three populations. For n scores in $[0, 1]$ with a mean of a, the one-sided Hoeffding lower bound is

$$L_\alpha = \max \left\{ 0,\ a - \sqrt{\frac{\log(1/\alpha)}{2n}} \right\}.$$

Applying $\alpha = 0.025$ for each ReMap condition to maintain a family-wise error rate of 0.05, the lower bounds calculate to 61.33 percent for two blocks and 59.91 percent for six blocks. Both figures significantly exceed the 10 percent cap, rejecting the hypothesis that every sampled population contains at most 40 distinct answer distributions.

C.3 Results at 400 Agents

Table 10 details every construction at 400 agents. Across the three Reddit seeds, matching yields standard deviations of 3.0, 2.5 and 2.9 points for reuse, two blocks and six blocks respectively. Reverse-direction matching achieves accuracies of 9.4, 65.5 and 63.6 percent. Removing answer length from the features prior to matching yields 11.0, 64.1 and 63.3 percent, whereas style matching produces accuracies of 9, 30 and 34 percent. For Congress, reverse matching sits at 8.75 percent for reuse and 60.25 percent for ReMap.

Assigning the ten copies of each Reddit module ten fixed temperatures ranging from 0.6 to 1.2 increases reuse matching to 19.5 percent, demonstrating a standard deviation of 0.66 points across seeds. The $1/\sqrt{3}$ mixture diverges noticeably from human text. On Reddit, 5.9 percent of its generated texts appear empty following quote removal. Text length alone separates these outputs from human text with scores of 67 to 80 across the three domains, whereas reuse and ReMap scores consistently stay between 48 and 59.

C.4 Matcher Range

Two reference checks demonstrate that the matcher successfully identifies distinct agents when they are present. When applied to the complete-module population using the true source as the label, it correctly identifies the 40 modules at a 21 percent rate utilizing just one answer per question. Supplying four and eight answers increases this accuracy to 87 and 98 percent, vastly outperforming the 2.5 percent chance level. When tested on 200 modules fitted separately to 200 distinct Reddit users, the matcher perfectly identifies every single agent. These reference modules rely on 80 posts each, share the identical training recipe and utilize their own author-disjoint human pool. Within that pool, their C2ST at 200 agents evaluates to 64.6, compared to 48.6 for two distinct human samples.

Table 10. Constructions at 400 agents from forty source modules on Qwen3-4B. C2ST, V and F use 800 free-writing texts.

Construction, 400 agents	**Matching** (% ↑)	**C2ST** (↓)	**V** (↑)	**F** (↓)	**Attribution** (× chance ↓)
Reddit *Means over three population seeds*					
Human reference		48.8	20.6	0.087	
Complete-module reuse	11.6	81.9	20.0	0.119	9.2
Reuse, varied temperature	19.5	81.7	19.9	0.123	10.1
ReMap, two blocks	65.2	81.4	19.7	0.121	3.5
ReMap, six blocks	63.8	83.3	19.3	0.126	1.6
Coefficient mixing, $1/3$	22.2	83.5	19.2	0.123	2.1
Coefficient mixing, $1/\sqrt{3}$	46.5	86.6	23.5	0.173	2.6
SCOTUS *One population seed*					
Human reference		47.7	21.4	0.091	
Complete-module reuse	8.5	68.2	17.4	0.105	19.0
ReMap, six blocks	55.2	79.1	15.5	0.119	1.8
Coefficient mixing, $1/\sqrt{3}$	35.5	77.8	18.1	0.140	3.6
Congress *One population seed*					
Human reference		50.3	27.1	0.099	
Complete-module reuse	9.5	78.6	21.3	0.150	14.7
ReMap, six blocks	60.5	85.9	20.6	0.155	1.6
Coefficient mixing, $1/\sqrt{3}$	44.5	89.7	17.1	0.168	3.3

D ReMap Analysis

D.1 Context Conditioning

Table 2 holds each agent's source individual and target lengths fixed, altering only the material the model sees. The representative condition selects the 16 posts closest to the semantic center of that individual's writing, while the final context condition introduces a persona distilled from the same archive. The SCOTUS rows merge data from three single-seed runs.

D.2 Module Ingredients

Table 11 details the comparison underlying Figure 5, incorporating Vendi and Fréchet distance metrics. The random update configuration redraws both LoRA factors from Gaussian distributions and rescales each adapted projection to match the Frobenius norm of the trained update. The shared one-source condition loads a single individual's module into every agent, whereas the shared all-source condition applies one rank-16 adapter trained on the combined text of all 40 individuals. The distinct conditions sample either complete modules or their individual blocks directly from the library.

We further investigate whether the shared all-source adapter merely requires additional training. The library originally trains 40 modules for 60 steps each at a batch size of 4, meaning 2,400 steps on the pooled text precisely match the total optimizer steps and the expected exposure rate of each post. We additionally save the adapter at 60 and 600 steps, subsequently training a rank-128 adapter for 2,400 steps using the identical recipe. Distinct modules achieve a lower C2ST than any shared adapter across every domain. The SCOTUS runs utilize the 120-post corpus from the regrouping comparison alongside its person-trained library.

The regrouping panel in Table 11 maintains 40 modules of 160 posts each and redistributes the identical 6,400 posts so that every module mixes text from k individuals. Each post appears exactly once, and every individual contributes an equal total number of posts. Single-source C2ST rises overall as each module incorporates more individuals, consistently remaining below the shared adapter baseline.

Table 11. Module and corpus-grouping comparisons on Qwen3-4B. Shared, all sources is one module fitted to all text with the library's total step count. Reddit grouping is a separate run.

Module comparison										
Configuration		**Reddit**			**SCOTUS**			**Congress**		
		C2ST	**V**	**F**	**C2ST**	**V**	**F**	**C2ST**	**V**	**F**
Human reference		52.0	14.0	0.32	52.0	15.2	0.33	51.0	18.0	0.36
Base		100.0	6.2	0.58	100.0	4.3	0.52	100.0	6.8	0.60
Random, single-source		100.0	6.7	0.56	100.0	5.0	0.51	100.0	7.4	0.58
Random, ReMap		100.0	6.9	0.56	100.0	4.9	0.52	100.0	7.3	0.58
Shared, one source		92.2	11.8	0.38	95.4	10.4	0.41	92.9	11.9	0.45
Shared, all sources		81.4	12.6	0.37	78.4	11.4	0.33	85.9	11.3	0.38
Distinct, single-source		76.4	14.1	0.34	67.5	13.2	0.34	74.0	14.7	0.38
Distinct, ReMap		82.2	13.1	0.33	77.2	11.7	0.34	81.9	13.7	0.38
Corpus grouping	*C2ST by people per module*									
Domain	**Configuration**	$k = 1$			$k = 4$			$k = 40$		**Shared**
Reddit	Single-source	76.3			76.0			78.3		81.4
	ReMap	82.4			80.7			80.9		n/a
Congress	Single-source	74.0			77.6			80.6		85.9
	ReMap	81.9			84.5			85.3		n/a
SCOTUS	Single-source	67.5			69.7			74.1		78.4
	ReMap	77.2			76.5			78.7		n/a

Table 12. Block count on Qwen3-4B, run separately under the protocol of Table 1. One block loads complete modules.

		Reddit			**SCOTUS**			**Congress**		
Blocks	**Recipes**	**C2ST** (↓)	**V** (↑)	**F** (↓)	**C2ST** (↓)	**V** (↑)	**F** (↓)	**C2ST** (↓)	**V** (↑)	**F** (↓)
Human		52.0	14.0	0.32	52.0	15.2	0.33	51.0	18.0	0.36
1	40	76.4	14.1	0.34	67.5	13.2	0.34	74.0	14.7	0.38
2	1,600	78.7	13.4	0.33	70.8	13.1	0.34	79.5	14.2	0.38
6	4.1 billion	82.2	13.1	0.33	77.2	11.7	0.34	81.9	13.7	0.38

D.3 Block Count

Table 12 compares setups utilizing one, two and six blocks under the protocol established in Table 1. This includes five repeats on Reddit and three on both SCOTUS and Congress for the two-block configuration. Moving from one to six blocks increases C2ST by 5.8 to 9.7 points and decreases Vendi by 1.0 to 1.5 points, all while the Fréchet distance remains stable.

Figure 8 illustrates the practical benefit of adding extra blocks. Using a single block, attribution to the contributing sources sits 4.9 to 8.5 above the permutation reference across all four settings. Increasing this to six blocks drops the margin to just 0.25 to 0.58 above the reference, and utilizing twelve blocks shrinks it to a maximum of 0.18. The Reddit sweep evaluating Qwen3-4B generates 40 unique recipes, producing twelve posts each at every evaluated block count.

The dependence sweep investigates whether blocks must be drawn independently. This test selects each subsequent source with a probability proportional to $\exp(\beta\kappa(i, j))$, where κ denotes the recorded behavioral similarity between two sources. This is evaluated on Reddit using Qwen3-4B with six blocks. Increasing β drastically reduces the number of active recipes by more than three orders of magnitude, maintains C2ST within 3.9 points of the independent draws and elevates source attribution at the highest setting. Consequently, independent draws serve as the default methodology.

D.4 Agent Coherence

ReMap posts successfully preserve sentence-level coherence. During the ReMap run utilizing Qwen3-4B, adjacent sentences within ReMap posts achieve cosine similarities of 0.60, 0.59 and 0.56 on Reddit, SCOTUS and Congress, against 0.56, 0.53 and 0.52 for human posts.

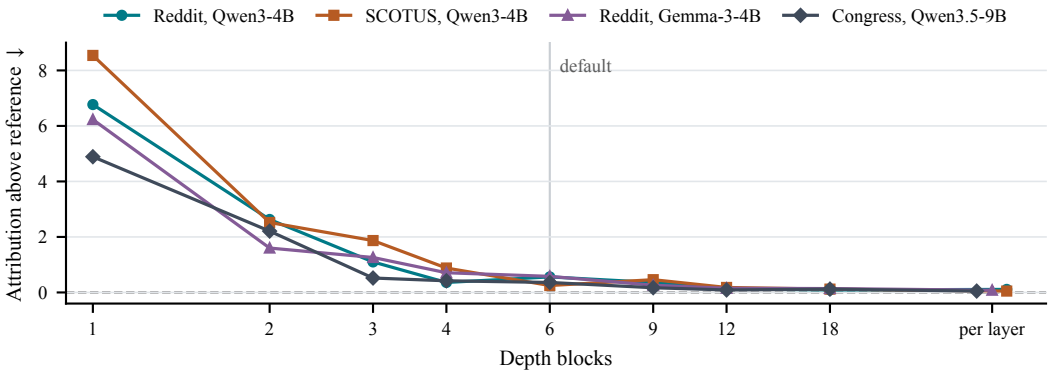

Figure 8. Source attribution above the permutation reference as the number of depth blocks grows. The last point of each line composes every layer separately.

Table 13. Training budget and regularization on Reddit. Near copies are percentages of generated posts.

Training	**Near copies** (↓)	**C2ST** (↓)
200 steps, no regularization	42	67.0
100 steps, dropout and weight decay	5	70.9
60 steps, dropout and weight decay	1	71.1

A specific recipe consistently reproduces the same agent across distinct generations. The Reddit expansion populations initialized with seeds 202 and 203 share 93 two-block recipes, identified strictly by the adapter content within each block. Treating one population as the gallery and the other as queries results in matching 80 and 77 out of the 93 recipes. This aligns closely with the 81 and 82 matches found internally within each population using those identical candidates. Style matching across the two populations yields scores of 51 and 55.

D.5 Text Reproduction

Table 13 demonstrates the necessity of training regularization. Near-copies of source posts are identified using a similarity threshold derived from pairs of human posts. For ReMap populations, every generated text across all domains remains below an embedding similarity of 0.92 when compared to the source texts. This verification checks each output against up to 600 source texts, sampling up to eight texts per source.

E Simulation Studies

E.1 Platform Populations

Figure 4 evaluates base and ReMap populations within the OASIS (Yang et al., 2024) and YSocial environments utilizing Qwen3-4B, Gemma-3-4B, Qwen3.5-9B and Gemma-3-12B. Each platform executes identical scenarios for Reddit and Congress, differing exclusively in their module libraries and human references. ReMap agents load one of 60 unique six-block recipes drawn from the 40 available modules within their respective domain, assigned deterministically by agent number.

Within OASIS, 111 agents share the neutral profile detailed in the prompt cards. The initial agent posts the vaccine-mandate claim, prompting all agents to post iteratively for twelve rounds while retaining their conversational memory. Agents are strictly limited to creating posts, requiring every turn to call the posting tool. Outputs are capped at 512 tokens, and each configuration executes a single simulation without applying a repetition penalty.

In YSocial, ReMap runs involve 80 agents operating over four simulated days, segmented into four time slots each. Agents post, comment, read and share with respective probabilities of 0.85, 0.4, 0.3 and 0.2, and topics are gun control, abortion, immigration and climate policy. Every agent maintains a distinct YSocial profile specifying age, gender, nationality, education, political leaning, Big Five personality traits and one to three interests. Agent memory is disabled, the news feed is empty and text sampling applies a temperature of 0.7 alongside a 200-token limit.

The resulting measurements exclusively analyze posts authored directly by the agents in OASIS and top-level posts in YSocial. Both sets of texts undergo the cleaning procedures outlined in Appendix A.1. C2ST calculations utilize 300 texts per class across every OASIS configuration and between 65 and 300 texts in YSocial, ensuring balanced matching within each setup. Vendi evaluations analyze the earliest available posts from each run up to those identical text counts.

E.2 OASIS STANCE

The six evaluated claims originate from ChangeMyView, addressing parenthood, gun policy, child support, rape culture, gender norms and religion. Each simulation run introduces one claim as the initial post. Both Qwen3-4B and Gemma-3-4B operate on the same vLLM build, utilizing a temperature of 0.7, a top-p of 0.8 and a top-k of 20 uniformly across all populations. Qwen3-4B additionally decodes with a repetition penalty of 1.2, a necessary adjustment because its chat template exposes each agent to its own earlier posts, a feature absent in the Gemma-3-4B template. Base agents share a single neutral profile, persona agents receive matched writing personas and ReMap agents load the composed modules from the platform study directly into their backbones. The human reference for each claim consists of its first 120 comments, strictly excluding any text that appears within the module training data.

DeepSeek-V4-Flash annotates every final-round post and human comment using an identical prompt. It classifies text along a five-point scale ranging from strong opposition to strong support, plus a label for posts that take no stance. Total variation distance compares these resulting six-category distributions, while a scored-only distance renormalizes just the five explicit stance categories. When including posts without a stance, the six-claim mean distances equal 0.47, 0.43 and 0.23 for base, persona and ReMap agents running on Qwen3-4B, alongside 0.49, 0.37 and 0.29 for those on Gemma-3-4B. The scored-only distances evaluate to 0.60, 0.41 and 0.19 on Qwen3-4B, compared to 0.46, 0.39 and 0.24 on Gemma-3-4B. During the final round, 26 to 49 percent of ReMap posts across either backbone repeat an earlier post from the same agent, a behavior rarely observed in base or persona agents. Excluding these repeated posts yields ReMap mean distances of 0.30 on Qwen3-4B and 0.23 on Gemma-3-4B, which still represent the lowest discrepancies among the three configurations.

E.3 OPINION INTERACTION

The vaccine-mandate study deploys 111 agents over twelve rounds, executing three independent repetitions. Agents receive the seed claim via the platform interface, retain their conversational memory, read the current feed prior to posting and execute actions using the provided platform tools. Every tested population decodes text with a repetition penalty of 1.2. We specifically analyze the original posts authored by the agents, discarding any reposts or direct quotes. Claude Haiku 4.5 (`claude-haiku-4-5-20251001`) evaluates each unique text exactly once using a sampling temperature of zero.

Figure 9 checks the Haiku labels against two human annotators across 200 posts sampled from earlier runs of this same study. These posts are drawn evenly from base, persona and single-source populations. On texts where the human annotators agree, they place 98, 85 and 66 percent of base, persona and single-source posts in the largest stance category, against 56, 61 and 50 percent for the judge. Consequently, the automated judge understates the ideological concentration of both the base and persona populations. Its overall agreement with the human annotators reaches $\kappa = 0.42$ across the three labels, compared to an inter-rater agreement of 0.66 between the humans themselves. Most discrepancies occur when the judge categorizes a post as neutral while the human annotators interpret it as explicitly opposed.

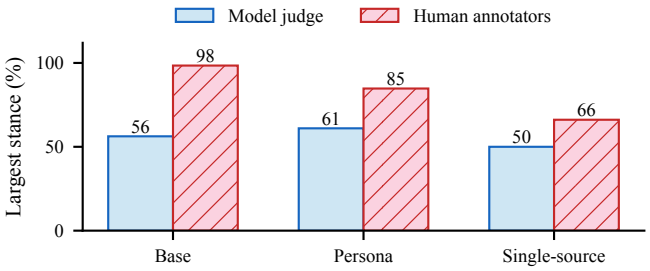

Figure 9. Share of posts in the largest stance category under the model judge and under human annotators, on the same posts.

To calculate stance entropy, we identify each agent's most frequent stance throughout the run and compute the Shannon entropy of these aggregated stances across all agents, measured in bits and averaged over the independent repeats. Jointly resampling agents across these repeats yields 95 percent confidence intervals of 0.22 to 0.38 for base agents, 0.55 to 0.78 for persona agents, 1.15 to 1.32 for single-source agents and 1.16 to 1.35 for ReMap agents. Conformity tracking focuses on explicit stance changes. A change qualifies as eligible when an agent authors a new text, its stance diverges from its previous position, the new stance registers as non-neutral and the preceding round featured a clear non-neutral majority. Conformity represents the proportion of these eligible changes that shift directly toward the prevailing majority.

The feed-swap test evaluates whether agents genuinely respond to their reading material. Two designated agents publish a pair of matched texts that consistently either endorse or reject the mandate, ensuring every other agent encounters precisely these two posts. Each agent subsequently performs one action, utilizing a sampling seed shared identically across both feed conditions. A third run replays the endorsing feed with these identical seeds to quantify how frequently labels fluctuate solely due to decoding noise. Single-source agents shift from 7 percent pro-mandate under the rejecting feed to 49 percent under the endorsing feed, while ReMap agents shift from 14 to 27 percent. Between the two feed conditions, 72 and 52 percent of their respective labels change, greatly exceeding the baseline noise levels of 12 and 14 percent. Neither base nor persona agents author any pro-mandate posts under either feed, and persona agents paradoxically display increased opposition after reading the endorsing content.

E.4 Information Responses

The information study executes separate false-claim and true-claim tasks within OASIS. This involves 121 agents acting over twelve rounds with three independent repeats under the previously described memory protocol. The false claims incorrectly link the MMR vaccine to autism, 5G networks to COVID-19 and vitamin C to COVID-19. Conversely, the efficacy of handwashing serves as the true claim. The human row in Table 3 provides the empirical share of survey respondents who hold these specific beliefs. This includes 10 percent of U.S. adults regarding vaccines and autism (Gallup, 2020), alongside 11 percent of Canadian residents for the claim that COVID-19 is spread to cover up harms of 5G (Carleton University, 2020). Claude Haiku 4.5 classifies each post as believing, debunking or uncertain based on a specific rubric for each claim. This classification operates at a temperature of zero, analyzing the first 800 characters of each text. Regarding handwashing, base, persona and ReMap posts accept the true claim at rates of 1.00, 0.15 and 0.39, while explicitly rejecting it at rates of 0.00, 0.38 and 0.09.

F Serving Cost

Table 4 benchmarks the full-model gather path using batches of 16 agents. The source sweep employs single-source routing on an NVIDIA GB10, processing 32 agents that generate 48 tokens each. The agent sweep constructs every agent from the 40 sources of a single domain on an H200 GPU, generating one length-matched post of up to 220 tokens per agent. Algorithm 1 details the gathered computation path for a single adapted projection. Given m active modules per agent, each

projection incurs a computational cost of $O(NTmrd)$. This contrasts favorably with $O(NTKrd)$ for the dense path, which must compute every module present in the library. Because a ReMap agent activates exactly one module per layer and its routing assignment updates exclusively when a new batch of agents begins generation, the gathered factors are efficiently reused across successive decoding steps. For extremely small libraries, the engine automatically switches to the dense path, which proves faster in those regimes while remaining numerically equivalent. Table 14 tracks modules stored in half precision within host memory, transfers the active ones to an H200 and isolates the cost of library management from the actual generation process. Each source module occupies approximately 34 MB in half precision. When utilizing 40 modules, throughput stabilizes at roughly 210 tokens per second from 32 agents onward, requiring a peak device memory footprint of about 11.2 GB. Finally, Table 15 compares the computational overhead of conditioning agents via a persona prompt, via 16 source posts provided in-context or via a single-source module.

Algorithm 1 Gathered composition for one adapted projection at layer ℓ.

Require: hidden states $X \in \mathbb{R}^{N\times T\times d}$ of N agents, frozen projection W, library factors $\{(A_k, B_k, s_k)\}_{k=1}^{K}$, routing indices $I_\ell \in \{1,\dots,K\}^{N\times m}$ and coefficients $C_\ell \in \mathbb{R}^{N\times m}$

$Y \leftarrow XW^\top$ ▷ frozen backbone
if I_ℓ changed since the previous decoding step **then** ▷ once per batch of agents
  $\hat{A}_{n,j} \leftarrow A_{I_\ell[n,j]}$ and $\hat{B}_{n,j} \leftarrow B_{I_\ell[n,j]}$ for all n, j ▷ gather active modules
  $f_{n,j} \leftarrow C_\ell[n,j]\, s_{I_\ell[n,j]}$ for all n, j
end if
$Z_{n,t,j} \leftarrow \hat{A}_{n,j} X_{n,t}$ for all n, t, j ▷ rank-r projection
$Y_{n,t} \leftarrow Y_{n,t} + \sum_{j=1}^{m} f_{n,j}\, \hat{B}_{n,j} Z_{n,t,j}$ for all n, t
return Y

Table 14. Host-memory paging for Qwen3-4B on an H200. Resident keeps every module on the GPU, and OOM marks a library larger than the 141 GB of device memory.

Sources	Resident GPU (GB)↓	Paged GPU (GB)↓	Host (GB)↓	Tokens/s↑
1,000	43	9.1	39	114
3,000	110	9.1	106	115
10,000	OOM	9.1	341	111

Table 15. Cost of conditioning 128 requests with up to 48 generated tokens each on Qwen3-4B. Relative gives latency against the persona prompt.

Conditioning	Tokens/s (↑)	ms per agent (↓)	Memory (GB) (↓)	Relative (↓)
Persona prompt	450	106.7	11.22	1.00
Sixteen posts in context	121	395.5	15.42	3.71
Single-source module	309	142.3	11.18	1.33