The Persona Fidelity Gap: Behaviorally Grounded LLM Personas Still Compress Real-User Preference Diversity

Rishav Kumar [1] **Atul Dev** [1] **Shivank Garg** [1]

Abstract

Large language models (LLMs) increasingly stand in for diverse human users in preference data collection, recommender evaluation, alignment research, and social-science simulation. Whether LLM personas reproduce real preference distributions has not been measured at scale, leaving downstream training, evaluation, and policy decisions exposed to silent bias. Prior persona work either skips behavioral validation or treats persona variability as a tunable hyperparameter without calibrating to a real distribution. We learn a preference embedding for roughly 268K Amazon-Reviews users via a multi-prototype Bradley-Terry encoder and place LLM personas in the same space using a Semantic Similarity Rating (SSR) protocol over stratified probe items. We compare grounded personas (seeded from a real user's history) and free-form personas against three classical synthetic baselines under a unified Fidelity Gap Index (FGI) over five distributional metrics. Grounding cuts FPD by roughly 30x over free-form prompting yet still leaves only 4.6% to 9.4% PRDC Coverage and intra-group cosine 0.94 to 0.99 against 0.56 to 0.68 for matched real users; item-level metrics look strong (rank accuracy 0.72 to 0.85) and hide this collapse. Local item-level fidelity does not buy global preference coverage, and the gap holds across two open-weights LLMs and four product splits. Our code is publicly available at `https://github.com/rixav77/persona-fidelity-gap`

1. Introduction

Large language models (LLMs) are increasingly cast as stand-ins for human users. They populate generative agent simulations (Park et al., 2023), fill in for survey respondents in computational social science (Argyle et al., 2023), supply preference labels for alignment research (Sorensen et al., 2024; Kirk et al., 2024a), and act as simulated raters for recommender and personalization systems. The implicit promise across these uses is the same: that a population of LLM personas behaves, in aggregate, like a population of real humans. That assumption holds only if the simulated population matches the real one in distribution rather than only on a handful of cherry-picked behaviors.

The cost of an undetected fidelity gap compounds across the pipelines that consume LLM persona populations. Reward models trained on simulated raters silently inherit any drift from real preferences, biasing alignment toward the central region of preference space and under-representing tails (Casper et al., 2023; Kirk et al., 2024b). Recommender simulators that A/B test new policies on synthetic users overstate gains when persona homogeneity collapses real user heterogeneity, and pluralistic-alignment evaluations that count the personas a model satisfies risk reporting coverage that does not exist (Sorensen et al., 2024; Kirk et al., 2024a). For reward modeling and pluralistic alignment specifically, this gap is load-bearing: a reward trained on a simulated population that systematically misses user tails will over-represent central preferences regardless of how accurately individual personas rank items. None of these concerns can be settled by inspecting individual persona behaviors; they require placing simulated and real populations in the same space and measuring the distance between the two distributions. We ask in this paper whether today's LLM personas honor the promise of population-level fidelity on observed user behavior, at scale.

Existing persona work does not answer this question. Some approaches do not validate against real behavioral data at all and instantiate populations from demographic templates or open-ended narrative prompts (Ge et al., 2024; Salewski et al., 2023). Others treat persona variability as a free hyperparameter tuned for downstream task accuracy rather than calibrated to a real distribution (Zollo et al., 2025). Prior work on diversity collapse under reinforcement learning from human feedback (Kirk et al., 2024b; Padmakumar & He, 2024; Santurkar et al., 2023) has documented a related

[1]IIT Roorkee. Correspondence to: Rishav Kumar <rishav_k1@ch.iitr.ac.in>, Atul Dev <atul_d@ma.iitr.ac.in>, Shivank Garg <shivank_g@mfs.iitr.ac.in>.

Pluralistic Alignment Workshop @ ICML 2026, Seoul, South Korea.

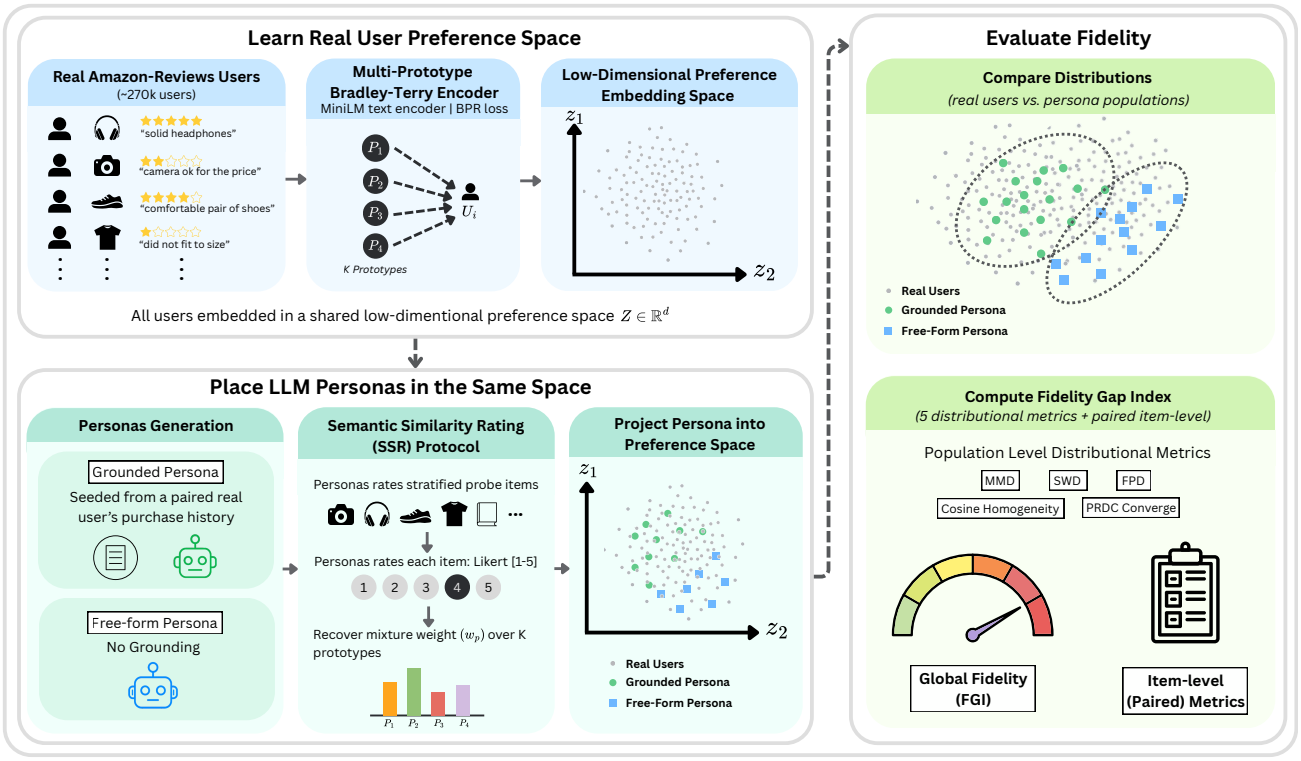

Figure 1. Pipeline overview. We learn a multi-prototype Bradley-Terry preference encoder over 268K real Amazon-Reviews users, place LLM personas in the same space via Semantic Similarity Rating (SSR), and compare grounded and free-form persona populations against held-out real users with five distributional metrics summarized through a Fidelity Gap Index (FGI).

effect, but the focus there is the diversity of generated text rather than the geometry of the preferences a persona population is supposed to represent. To our knowledge, no existing protocol places LLM personas and real users in a learned preference space and measures the distance between the two distributions at scale.

We use the term *fidelity gap* for that distance: the discrepancy between the empirical distribution induced by a population of LLM personas and the empirical distribution of real users in a learned preference space. The embedding must be learned from observed behavior rather than borrowed from a generic text encoder, because text similarity is not preference similarity, and behaviorally grounded personas are where the open question lies: free-form templates have already been argued to be too unconstrained to recover any specific human (Ge et al., 2024), so the question is whether grounding on a user's actual history is sufficient to recover that user's distributional neighborhood.

Our approach is summarized in Figure 1. We learn a multi-prototype Bradley-Terry encoder (Chen et al., 2025) on roughly 268K real Amazon-Reviews users (Hou et al., 2024), giving each user a low-dimensional preference vector from a frozen MiniLM (Reimers & Gurevych, 2019; Hugging Face & Sentence-Transformers, 2021) item context and a learnable mixture over shared prototype projectors. We then place LLM personas in the same space using a Semantic Similarity Rating (SSR) protocol: each persona rates stratified probe items on a Likert scale, and we recover the per-persona mixture weight by optimizing the same pairwise Bayesian Personalized Ranking (BPR) objective (Rendle et al., 2009) that trained the encoder, with projectors frozen. We evaluate three persona conditions (free-form demographic templates, free-form narrative backstories, behaviorally grounded personas seeded from a paired real user's top purchases) against held-out real users and three classical synthetic-user baselines (Dirichlet, Gaussian mixture, prototype perturbation). We score every population on five distributional metrics (Frechet Preference Distance, Maximum Mean Discrepancy, Sliced Wasserstein, PRDC Coverage, intra-group cosine homogeneity) summarized through a unified Fidelity Gap Index, and on item-level paired metrics for the grounded condition.

Across four product splits and two open-weights LLMs (Llama-3.1-8B and Qwen3-4B), we find a consistent pattern. Behavioral grounding helps at the item level, where grounded personas reach paired rank accuracy 0.72 to 0.85 and per-user rating correlations up to 0.84. At the population level, the same personas recover only 4.6% to 9.4% of real-user neighborhoods (PRDC Coverage) and collapse to

intra-group cosine 0.94 to 0.99 versus 0.56 to 0.68 for the matched real users. Table 1 and Figure 2 summarize headline numbers per split and the per-metric gap normalized to a real-bootstrap floor and a moment-matched random ceiling. The picture is the same for both LLMs we test, so the gap is a property of the prompting and rating protocol rather than of any specific base model. Our main contributions include:

- We propose a learned-embedding fidelity gap measurement protocol that places LLM personas and real users in the same low-dimensional preference space, validated at scale on 268K Amazon-Reviews users across four product splits via a frozen-encoder SSR procedure.
- We define the Fidelity Gap Index (FGI), a unified normalization that places five distributional metrics on a common 0–1 scale anchored at a real-bootstrap floor and a moment-matched random baseline, so different metrics, splits, and persona sources can be compared directly.
- We provide empirical evidence that behaviorally grounded LLM personas achieve high item-level rank fidelity yet collapse in population coverage and homogeneity; a local-global tradeoff missed by prior persona evaluations focused on per-user accuracy.
- We release a reproducible pipeline (preference encoder, three-condition persona generation, SSR rating, paired and distributional evaluation) into which other persona generators or LLMs can be substituted.

2. Related Work

Pluralistic alignment and persona simulation. Single-reward alignment objectives flatten human heterogeneity, motivating pluralistic alternatives. Sorensen et al. (2024) lay out a roadmap for representing many viewpoints in a single model, and Bakker et al. (2022) fine-tune language models to find agreement among raters with diverse preferences. Kirk et al. (2024a) contribute the PRISM dataset of multicultural and subjective human feedback; PRISM offers ground-truth diversity but does not learn a continuous preference space in which to place LLM personas. On the simulation side, Argyle et al. (2023) use LLMs as silicon survey respondents, Park et al. (2023) embed LLM agents in interactive simulacra, and Salewski et al. (2023) probe in-context impersonation. Persona populations in this line are evaluated by qualitative inspection or downstream accuracy rather than by distance to an observed-behavior distribution.

Persona generation and benchmarking. A complementary literature scales persona populations or uses them as preference proxies. Ge et al. (2024) construct a persona hub with about a billion templated identities, Lee et al. (2024) align language models to thousands of preferences via system message generalization, and Santurkar et al. (2023) measure whose opinions current LLMs reflect. Closest to us, Zollo et al. (2025) introduce PersonalLLM, where a learned reward model *induces* synthetic users; the reward model substitutes for users rather than anchoring personas to observed behavior, and the evaluation targets per-user accuracy. Recommender simulation work using LLM-driven users for cold-start and counterfactual evaluation inherits the same risk: a geometrically narrower synthetic population systematically misestimates tail performance.

Diversity collapse and distributional metrics. A separate line documents diversity loss in language model *outputs*: Kirk et al. (2024b) report that RLHF narrows generation diversity, Padmakumar & He (2024) show that writing with an LLM reduces content diversity, and Casper et al. (2023) survey RLHF limitations. Our object is different: geometric collapse in a learned *preference space*, not lexical collapse in text, and the two need not coincide. To quantify population fidelity we adapt the Frechet distance (Heusel et al., 2017), MMD (Gretton et al., 2012), sliced Wasserstein (Bonneel et al., 2015), and the precision/recall/density/coverage family (Kynkäänniemi et al., 2019; Naeem et al., 2020), unified through a single Fidelity Gap Index.

Synthesis. The SSR-on-PAL placement protocol fills the absence of any method that grounds LLM personas in a learned space built from observed user behavior (Chen et al., 2025; Kirk et al., 2024a; Zollo et al., 2025), going beyond reward-model surrogates and templated identities. The Fidelity Gap Index brings the distributional view that prior diversity-collapse studies applied to text (Kirk et al., 2024b; Padmakumar & He, 2024) into preference geometry, aggregating metrics so far used in isolation. The local-global divergence we report (high item-level fidelity, collapsed population coverage) is invisible to per-user accuracy benchmarks (Zollo et al., 2025; Salemi et al., 2024) by construction, with direct consequences for any pipeline relying on simulated users for offline evaluation. The pipeline release packages all four steps so that subsequent pluralistic-alignment work can be measured against a real-user reference distribution rather than a synthetic surrogate.

3. Method

We learn a low-dimensional preference embedding for real users, generate three families of LLM personas, place those personas into the same space using a Semantic Similarity Rating (SSR) protocol, and compare the populations through five distributional metrics and a unified Fidelity Gap Index (FGI). The encoder is held frozen during SSR by design:

this cleanly attributes any measured gap to the persona-prompting protocol rather than to encoder adaptation, while leaving open that protocols which jointly fine-tune both could narrow it.

3.1. Preference embedding

Let $\mathcal{U} = \{1, \ldots, N\}$ index N real users. We adapt the multi-prototype Bradley-Terry encoder of Chen et al. (2025) to user-item review data, training the user representation by Bayesian Personalized Ranking (BPR) over pairwise rating gaps (Bradley & Terry, 1952; Rendle et al., 2009). The backbone is a frozen sentence encoder $\phi : \text{text} \to \mathbb{R}^{384}$, instantiated as all-MiniLM-L6-v2 (Reimers & Gurevych, 2019; Hugging Face & Sentence-Transformers, 2021).

Each user i carries learnable mixture weights $\mathbf{w}_i \in \mathbb{R}^K$, and the model shares K MLP projectors $P_k : \mathbb{R}^{384} \to \mathbb{R}^d$ across users. Given a context vector $\mathbf{c} = \phi(\text{text}) \in \mathbb{R}^{384}$, the user representation is the convex combination

$$u_i(\mathbf{c}) = \sum_{k=1}^{K} \pi_{ik}\, P_k(\mathbf{c}), \qquad \boldsymbol{\pi}_i = \mathrm{softmax}(\mathbf{w}_i/\tau), \quad (1)$$

with mixture temperature τ. Items are scored using the same projector family applied to the item text t, with no per-item parameters. With normalization $\hat{\mathbf{v}} = \mathbf{v}/\|\mathbf{v}\|$, the score of item t for user i is

$$s(i,t) = \left\langle \widehat{u_i(\mathbf{c})},\ \widehat{u_i(\phi(t))} \right\rangle e^{\sigma}, \quad (2)$$

where σ is a learnable scalar (CLIP-style logit temperature). Both context and item are routed through the same user-specific mixture, so the same item appears differently to different users, which is the multi-prototype design of Chen et al. (2025). For any pair (a, b) rated by user i with $r(i,a) - r(i,b) \geq 1$, training minimizes

$$\mathcal{L}_{\text{CE}} = -\log \frac{e^{s(i,a)}}{e^{s(i,a)} + e^{s(i,b)}}. \quad (3)$$

We train with $K = 4$ following the PAL ablations of Chen et al. (2025), embedding dimension $d = 128$, OneCycleLR, fp16 mixed precision, batch size 512, and 20 epochs, validating by held-out pairwise accuracy; full configuration in the appendix.

After training, real-user embeddings are extracted at a single reference context. Let $\bar{\mathbf{c}} = \frac{1}{|\mathcal{I}|} \sum_{t \in \mathcal{I}} \phi(t)$ be the mean over the item catalog $\mathcal{I}$; the frozen real-user embedding is

$$u_i^* = u_i(\bar{\mathbf{c}}) \in \mathbb{R}^d. \quad (4)$$

With the context fixed, u_i^* depends only on $\boldsymbol{\pi}_i$ and the shared projectors, so any two users differ only through their mixture weights. Fixing the context to the catalog mean decouples user comparison from item conditioning, so distributional metrics over $\{u_i^*\}$ measure preference-geometry differences alone; alternatives such as a per-user item-mean would entangle user differences with which items each user happened to consume.

3.2. Three persona conditions

We compare three persona-generation strategies that share the downstream pipeline. We draw on earlier persona-prompting recipes (Ge et al., 2024; Zollo et al., 2025) and isolate the empirical question of how much each strategy preserves real preference structure. In every condition an open-weights chat model is prompted to emit a free-text persona, which is then embedded by SSR (Section 3.3).

Free-form Demographic prompting. The LLM is given hard demographic attributes (age, gender, location, occupation, income, education) sampled from realistic marginal distributions, and is asked to write a 150 to 250 word persona. Decoding uses temperature T=0.9 and top-p=0.95. No item-level information enters the prompt.

Free-form Backstory prompting. The LLM is asked for a 200 to 300 word open-ended narrative under the same sampler. There are no attribute slots and no item history; the persona is constrained only by stylistic instructions.

Behaviorally Grounded prompting. A real user's top-20 highest-rated items are formatted into a textual profile with rating, category tag, truncated title, and truncated review snippet, alongside the empirical rating distribution. The LLM is asked for a 150 to 250 word persona that stays grounded in the evidence and does not invent preferences that contradict the listed history. We preserve the seed user-id, so every grounded persona is paired with the real user it was conditioned on; this enables paired evaluation downstream.

The methodological contrast is that conditions (a) and (b) make no contact with observed behavior, while condition (c) does. Any fidelity difference between (c) and the free-form conditions is therefore attributable to behavioral grounding alone.

3.3. Persona embedding via Semantic Similarity Rating

To place each persona in the same $\mathbb{R}^d$ as real users, we recover its mixture weight by treating the persona as a black-box rater of probe items. We fix a probe set $\mathcal{P}$ of $M = 100$ items drawn by stratified k-means clustering with three popularity strata (head, torso, tail by review count) at ratio 20/40/40, so the probes cover both common and niche regions of the catalog. The probe count $M = 100$ was selected during pilot validation to give adequate pair coverage at modest API cost.

Each persona is asked, for every probe $p \in \mathcal{P}$, "how likely would you be to purchase this product on a 1 to 5 scale?", at temperature 0.1 to suppress sampling noise. From the resulting Likert ratings $\{r_p\}_{p\in\mathcal{P}}$ we extract all pairs (a, b) with $|r_a - r_b| \geq 1$, which matches the encoder's pair-mining rule. With $M = 100$ this yields up to $\binom{100}{2} = 4950$ candidate pairs; in practice roughly 60–80% survive the rating-gap filter (see appendix), so the per-persona BPR fit below is well-posed rather than under-determined.

We then optimize the persona's mixture weight $\mathbf{w} \in \mathbb{R}^K$ on the same BPR-style cross-entropy loss the encoder used, with both the projectors $\{P_k\}$ and the sentence-embedding backbone ϕ held fixed. With $u_{\mathbf{w}}(\bar{\mathbf{c}}) = \sum_k \mathrm{softmax}(\mathbf{w}/\tau)_k P_k(\bar{\mathbf{c}})$ and $s_{\mathbf{w}}(t) = \langle \widehat{u_{\mathbf{w}}(\bar{\mathbf{c}})}, \widehat{u_{\mathbf{w}}(\phi(t))} \rangle e^{\sigma}$, the persona's mixture weight is

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \sum_{(a,b)} -\log \mathrm{sig}\big(s_{\mathbf{w}}(a) - s_{\mathbf{w}}(b)\big), \qquad (5)$$

where sig is the logistic function. We solve this with 200 steps of Adam at learning rate 0.01. The final persona embedding is $u_{\hat{\mathbf{w}}}(\bar{\mathbf{c}}) \in \mathbb{R}^d$, in the same space and at the same reference context as the frozen real-user embeddings u_i^*, so the two populations are directly comparable. Pseudocode is in Algorithm 1.

3.4. Distributional fidelity metrics and FGI

Given two equal-size sets of d-dimensional embeddings X (real subsample) and Y (synthetic or LLM personas), we report five distributional metrics.

- **Frechet Preference Distance (FPD)** (Heusel et al., 2017; Kynkäänniemi et al., 2019): the Frechet distance between Gaussian fits, $\mathrm{FPD}(X, Y) = \|\boldsymbol{\mu}_X - \boldsymbol{\mu}_Y\|^2 + \mathrm{Tr}\big(\boldsymbol{\Sigma}_X + \boldsymbol{\Sigma}_Y - 2(\boldsymbol{\Sigma}_X\boldsymbol{\Sigma}_Y)^{1/2}\big)$.
- **Maximum Mean Discrepancy (MMD)**: the unbiased squared MMD with a Gaussian kernel at a median-heuristic bandwidth (Gretton et al., 2012), with a permutation p-value at $B = 500$.
- **Sliced Wasserstein Distance (SWD)** (Bonneel et al., 2015): the average 1-D Wasserstein distance over $L = 500$ random unit-norm projections.
- **PRDC Coverage at** $k = 5$ (Naeem et al., 2020): the fraction of real points whose k-nearest-neighbor ball contains at least one synthetic point.
- **Intra-group cosine homogeneity**: the mean pairwise cosine on a 500-row sample of Y, used as a tightness diagnostic for collapse onto a centroid.

These five metrics are not on the same scale, and a model can do well on one and poorly on another. We therefore rescale each through a shared reference frame, the *Fidelity Gap Index* (FGI).

Let $\mathcal{R}_{\text{boot}}$ be a held-out 500-row bootstrap of real users (an empirical floor), and $\mathcal{R}_{\text{rand}}$ a moment-matched Gaussian baseline at the real distribution's mean and per-coordinate (diagonal-covariance) standard deviation, serving as an upper bound for "no preference structure". The diagonal choice is deliberate: a full-covariance Gaussian would absorb the prototype-mixture correlations that distinguish real users and bury the FGI signal, whereas a diagonal anchor preserves only the marginal moments. For metrics where lower is closer to real (FPD, MMD, SWD), we set

$$\mathrm{FGI}_m(Y) = \frac{m(Y) - m(\mathcal{R}_{\text{boot}})}{m(\mathcal{R}_{\text{rand}}) - m(\mathcal{R}_{\text{boot}})}. \qquad (6)$$

For Coverage, where higher is closer to real, the numerator and denominator are reversed:

$$\mathrm{FGI}_{\text{Cov}}(Y) = \frac{\mathrm{Cov}(\mathcal{R}_{\text{boot}}) - \mathrm{Cov}(Y)}{\mathrm{Cov}(\mathcal{R}_{\text{boot}}) - \mathrm{Cov}(\mathcal{R}_{\text{rand}})}. \qquad (7)$$

Interpretation is uniform: $\mathrm{FGI} = 0$ means Y is indistinguishable from a real bootstrap on metric m, $\mathrm{FGI} = 1$ means Y is indistinguishable from random on m, and $\mathrm{FGI} > 1$ means Y is worse than moment-matched noise on that metric.

Paired metrics for grounded personas. Because grounded personas preserve the seed user-id, we report additional statistics on the one-to-one matching between persona and seed: per-row paired cosine and L2 distance, intra-group cosine on the matched real subset, and a paired rank-accuracy score on 10K random (user, item, item) triplets. We also compute a per-user Pearson rating correlation across 100 random items between persona Likert ratings and the encoder-predicted ratings of the seed user u_{i^*} scored on the same probes; we use the seed's encoder predictions because the held-out probe set has no observed real ratings. These paired diagnostics probe item-level fidelity that the population-level metrics above cannot.

4. Experimental Setup

Datasets. We evaluate on the Amazon Reviews 2023 corpus (Hou et al., 2024; Ni et al., 2019) using the 5-core extension. We construct four splits: Musical_Instruments (50,684 users, 16,844 items), Software (136,780 users), Video_Games (88,779 users), and Combined (267,967 users), the concatenation. Per-user interactions are ordered by timestamp and partitioned 80/10/10 chronologically into encoder training, validation, and held-out evaluation pools. Per-split mean / std of the real intra-user pairwise cosine are 0.715 / 0.237 (Combined), 0.752 / 0.245 (Musical_Instruments), 0.680 / 0.257 (Software), 0.665 / 0.274 (Video_Games).

Models. We test two open-weights instruction-tuned LLMs: meta-llama/Llama-3.1-8B-Instruct (Dubey et al., 2024) and Qwen/Qwen3-4B-Instruct-2507 (Yang et al., 2025), both served via vLLM (Kwon et al., 2023) on a single 24 GB GPU with FP16 weights. Persona generation uses temperature 0.9, top-p 0.95, 512-token cap; SSR item rating uses temperature 0.1, top-p 0.95, 8-token cap, since only integers in $\{1, 2, 3, 4, 5\}$ are valid. Both LLMs share identical prompts and decoding parameters.

Synthetic baselines. To separate the LLM-persona fidelity gap from generic distribution-matching difficulty, we include three classical samplers that operate on the same $\mathbb{R}^d$ preference space as real users.

- *Dirichlet interpolation*: $\mathbf{u}_{\text{syn}} = \sum_{j=1}^{n_{\text{anchor}}} \alpha_j \mathbf{u}_{i_j}$ with $\boldsymbol{\alpha} \sim \text{Dir}(\beta = 1)$ and $n_{\text{anchor}} = 3$. Anchors are drawn uniformly from real users.
- *Gaussian Mixture sampler*: a 16-component diagonal-covariance GMM fit to the real-user embedding cloud and then resampled.
- *Prototype perturbation*: a Dirichlet-sampled mixture over the K learned prototypes, plus isotropic Gaussian noise with $\sigma = 0.01$.

None of the three uses the LLM, so they isolate how much of the gap comes from population geometry versus the persona-prompting and SSR pipeline.

Implementation details. We sample 500 grounded personas per (split, model) pair, each paired with a real user with at least 10 reviews. Each persona rates 100 stratified probe items (head/torso/tail by popularity, k-means diversified inside each band); we recover the persona embedding via 200 Adam steps on BPR with encoder projectors frozen. Real-user fidelity baselines come from 500 bootstrap resamples of held-out users; the moment-matched random Gaussian baseline draws coordinates from $\mathcal{N}(\mu_d, \sigma_d^2)$. Distributional metrics are mean over five resamples of size 500; MMD permutation test uses $B = 500$. The encoder uses $K = 4$ prototypes and $d = 128$. Hyperparameters are in Section A.

5. Results

5.1. A large persona fidelity gap across splits

Grounded LLM personas miss the real-user distribution by a wide margin in every cell. PRDC Coverage ranges from 4.6% (Video_Games / Qwen) to 9.4% (Combined / Qwen), roughly an order of magnitude below the 88.4% to 96.0% from Dirichlet interpolation (which achieves high coverage by construction) and the 93.2% to 98.6% from a held-out real bootstrap. In normalized terms (Table 1), FGI-Coverage sits at 0.91 to 0.95, near the random ceiling of 1.0; FGI-MMD often exceeds 1.0 because the LLM cloud is tighter than the random Gaussian and farther from the real moments in kernel space. The prototype baseline also drops to zero coverage, but for a different reason: it lives on a low-rank K-simplex (Table 2). Both LLMs trace nearly identical curves in Figure 2, so the gap reflects the protocol, not any one base model.

5.2. Diversity collapse in intra-group cosine

LLM personas cluster much more tightly than the matched seeds: mean intra-group cosine across the eight (split, model) cells is 0.94 to 0.99, against 0.56 to 0.68 for matched real users. Intra-cosine standard deviation falls from 0.24 to 0.31 (real) to 0.04 to 0.07 (LLM). Figure 3 plots both side by side; real users span the manifold while LLM personas pile up near a single mode regardless of split (Figure 5 in the appendix shows the 2D overlay). The matched real intra-cosine tracks the unconditional split-level statistic, so the seed pool is representative.

5.3. Local-global divergence

The same grounded personas that fail at population coverage do well on item-level paired metrics: paired rank accuracy on 10K (user, item, item) triplets is 0.72 to 0.85, and per-user Pearson correlation between the persona's 100 probe ratings and the seed user's reconstructed ratings is 0.57 to 0.84. These are conditional scores; the LLM sees the seed's history and only needs to recover relative orderings. Figure 4 plots the tradeoff: Software has the strongest local fidelity but Coverage matches Musical_Instruments and Video_Games. We call this a *local-global divergence*: high item-level fidelity does not imply distributional faithfulness. Held-in coverage of the full real population is 0.0034 to 0.0076, so the gap is not a sampling artifact. Per-row paired cosine distance concentrates within 0.3 of the seed, so local fidelity is genuine (Figure 6).

5.4. Grounding helps relative to free-form, but the gap persists

Behavioral grounding improves over free-form prompting but does not close the gap. On Musical_Instruments with Llama-3.1-8B, free-form demographic and backstory personas land at Coverage 0.000 with intra-LLM cosine 0.981 and 0.989; grounded personas reach Coverage 0.084 and FPD 4.4×10^{-4}, an order of magnitude below the 1.4×10^{-2} FPD of free-form (Table 3). Grounding buys global signal: the cloud's first two moments move toward real, and Coverage moves off zero. The residual gap is still substantial, with FGI-MMD 11 to 21 times the random baseline and FGI-SWD about 1.2 times. Grounding is necessary but not

Table 1. Per-cell fidelity metrics for n=500 grounded personas across four Amazon-Reviews splits and two open-weights LLMs. Cov: PRDC Coverage (proportion in $[0, 1]$); Cos LLM: intra-group cosine; Pair cos: paired cosine to seed; Rank acc: rank accuracy on 10K triplets; Rate corr: per-user Pearson rating correlation; FGI-Cov, FGI-MMD: Fidelity Gap Index for Coverage and MMD. Matched intra-real cosine is 0.56–0.68 across splits; full numbers in the appendix.

SPLIT	MODEL	COV	COS LLM	PAIR COST-DIST	RANK ACC	RATE CORR	FGI-COV	FGI-MMD
COMBINED	LLAMA-3.1-8B	0.088	0.943	0.227	0.797	0.787	0.911	7.39
COMBINED	QWEN3-4B	0.094	0.964	0.214	0.805	0.796	0.905	8.84
MUSICAL_INSTR	LLAMA-3.1-8B	0.084	0.985	0.269	0.729	0.579	0.914	8.71
MUSICAL_INSTR	QWEN3-4B	0.066	0.986	0.274	0.722	0.566	0.933	9.55
SOFTWARE	LLAMA-3.1-8B	0.088	0.977	0.249	0.843	0.839	0.908	13.16
SOFTWARE	QWEN3-4B	0.078	0.978	0.241	0.846	0.841	0.919	12.81
VIDEO_GAMES	LLAMA-3.1-8B	0.060	0.983	0.272	0.805	0.754	0.936	10.69
VIDEO_GAMES	QWEN3-4B	0.046	0.988	0.278	0.803	0.752	0.951	12.50

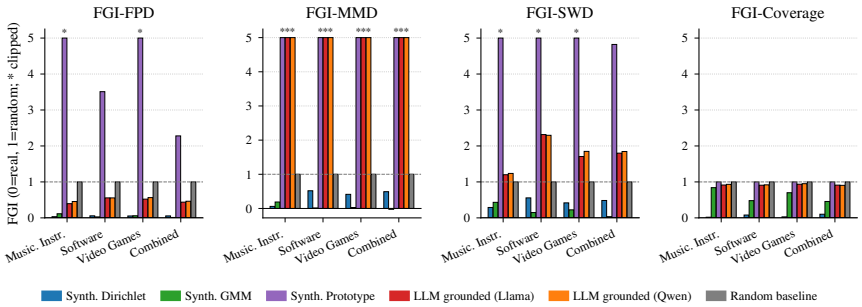

Figure 2. Fidelity Gap Index (FGI) per metric across the four splits and the two LLMs. FGI = 0 is indistinguishable from a real bootstrap; FGI = 1 is as bad as a moment-matched random Gaussian. On Coverage, grounded personas sit at 0.91 to 0.95, near the random ceiling, while Dirichlet interpolation lies near 0.10. On MMD and SWD, LLM personas exceed the random baseline (FGI > 1).

Table 2. Distributional metrics on the Combined split, n=500 per row. FPD, MMD, and SWD are reported in units of 10^{-3}; Cov and Cos are proportions in $[0, 1]$. Lower is better for FPD, MMD, SWD; higher for Coverage. The random baseline is moment-matched diagonal Gaussian noise. MMD permutation p-values are 0.323 (real_subsample) and 0.671 (synthetic_gmm); all other rows have $p < 0.005$.

METHOD	FPD	MMD	COV	SWD	COS
REAL_SUBSAMPLE	0.084	0.27	0.986	0.55	0.699
SYNTHETIC_DIRICHLET	0.45	18	0.886	1.3	0.819
SYNTHETIC_GMM	0.16	−0.63	0.536	0.60	0.713
SYNTHETIC_PROTOTYPE	15	220	0.000	7.8	0.233
LLM_GROUNDED (LLAMA)	3.0	272	0.088	3.3	0.943
LLM_GROUNDED (QWEN)	3.2	326	0.094	3.3	0.964
RANDOM_BASELINE	6.8	37	0.000	2.1	0.639

sufficient.

5.5. Norm collapse: LLM personas under-represent the manifold radius

The L2-norm ratio between LLM personas and real users sits at 0.69 to 0.82 across all eight cells, so persona embeddings are systematically pulled toward the centroid (Figure 7).

Table 3. Effect of grounding on Musical_Instruments / Llama-3.1-8B. Free-form personas ($n = 100$) collapse to zero coverage; grounding ($n = 500$) cuts FPD by roughly 30x and lifts Coverage to 8.4%.

METHOD	FPD	MMD	COV	COS
LLM_DEMOGRAPHIC	1.4E-2	1.42	0.000	0.981
LLM_BACKSTORY	1.3E-2	1.48	0.000	0.989
LLM_GROUNDED	4.4E-4	0.421	0.084	0.985
RANDOM_BASELINE	6.8E-3	3.7E-2	0.000	0.639

Combined with high intra-cosine (Section 5.2), personas occupy a small near-spherical cluster around the mean. This explains why grounded personas pass per-user checks while failing population checks: per-row geometry is recovered, but the dispersion defining the real population is not.

5.6. Cross-LLM consistency

Llama-3.1-8B-Instruct and Qwen3-4B-Instruct-2507 produce nearly indistinguishable fidelity gaps. Per-cell Coverage differs by 0.012 to 0.018, intra-LLM cosine by less than 0.02. Both LLMs trace the same shape in Figure 2 and the same ranking in Table 1, so the gap is a property of

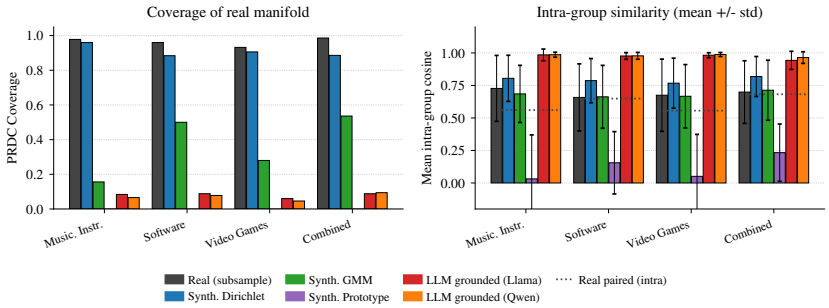

Figure 3. Left: PRDC Coverage of the real preference distribution by each source. Right: mean intra-group cosine for the same source. Grounded LLM personas sit roughly an order of magnitude below the Dirichlet baseline on Coverage and cluster at 0.94 to 0.99 mean cosine, well above the 0.56 to 0.68 of matched real users.

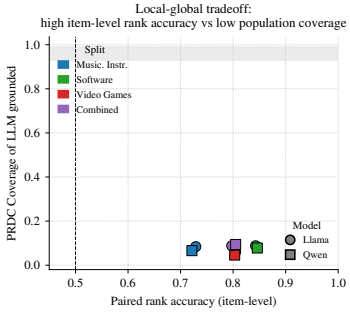

Figure 4. Per-cell scatter of paired rank accuracy (item-level) against PRDC Coverage (population-level). The separation between the two axes is the local-global divergence: grounded personas can recover the seed user's preference ordering reasonably well while still failing to populate the rest of the real-user manifold.

the persona-prompting plus SSR protocol, not of any one model's stylistic priors.

6. Discussion

6.1. Why the gap persists despite grounding

Behavioral grounding feeds the LLM a prose summary of a user's top-rated items, then asks it to re-rate stratified probes at T=0.1. Two losses compose: the rendered profile is a lossy summary, dropping rare items from the top-20 window; and the LLM's rating function inherits modal regularities that low-temperature SSR compresses further. Text-diversity collapse under RLHF and many-shot prompting is documented (Padmakumar & He, 2024; Kirk et al., 2024b); we extend it to preference-geometry collapse. The high paired rank accuracy (Table 1) confirms the LLM reads the seed user faithfully at the item level; population shrinkage is a joint consequence of the text bottleneck and mode-seeking rating function.

6.2. Implications for persona-based alignment and evaluation

Pluralistic-alignment work (Sorensen et al., 2024; Kirk et al., 2024a) and persona-based reward modeling (Zollo et al., 2025; Lee et al., 2024; Chen et al., 2025) take persona variability for granted. Even with grounding, populations occupy a thin slice of real preference space (intra-cosine 0.94 to 0.99 vs 0.56 to 0.68), reaching only 4.6% to 9.4% PRDC Coverage. Pipelines training on simulated populations risk over-fitting the central region and under-representing tails (Casper et al., 2023). Item-level metrics mask this by rewarding local agreement without penalizing global homogeneity; evaluations should report distributional fidelity alongside per-user accuracy, with FGI as a practical anchor.

7. Conclusion

We present a learned-embedding protocol for measuring how faithfully LLM persona populations reproduce a real-user preference distribution, a unified Fidelity Gap Index that places five distributional metrics on a common scale, and an empirical study against 268K Amazon-Reviews users. Across four product splits and two open-weights LLMs, behaviorally grounded personas reach paired rank accuracy 0.72 to 0.85 yet cover only 4.6% to 9.4% of real-user neighborhoods and collapse to intra-group cosine 0.94 to 0.99 against a matched real range of 0.56 to 0.68. The disparity between item-level fidelity and population coverage is consistent across both LLMs and all four splits, which indicates a property of the persona-prompting and SSR protocol rather than of any specific base model; item-level accuracy is therefore not a sufficient proxy for the distributional fidelity that downstream pipelines depend on. Closing the gap is an open question, with protocols that jointly fine-tune the encoder, condition on richer behavioral context than a top-20 summary, or explicitly diversify decoding as natural next steps. We release the full pipeline as an FGI-anchored benchmark and recommend that persona-based alignment,

recommender, and social-science evaluations report distributional fidelity alongside item-level accuracy.

References

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., and Wingate, D. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31 (3):337–351, 2023.

Bakker, M. A., Chadwick, M. J., Sheahan, H. R., Tessler, M. H., Campbell-Gillingham, L., Balaguer, J., McAleese, N., Glaese, A., Aslanides, J., Botvinick, M. M., and Summerfield, C. Fine-tuning language models to find agreement among humans with diverse preferences. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Bonneel, N., Rabin, J., Peyré, G., and Pfister, H. Sliced and Radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015.

Bradley, R. A. and Terry, M. E. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., Wang, T., Marks, S., Segerie, C.-R., Carroll, M., Peng, A., Christoffersen, P., Damani, M., Slocum, S., Anwar, U., Siththaranjan, A., Nadeau, M., Michaud, E. J., Pfau, J., Krasheninnikov, D., Chen, X., Langosco, L., Hase, P., Bıyık, E., Dragan, A., Krueger, D., Sadigh, D., and Hadfield-Menell, D. Open problems and fundamental limitations of reinforcement learning from human feedback. *Transactions on Machine Learning Research (TMLR)*, 2023.

Chen, D., Chen, Y., Rege, A., Wang, Z., and Vinayak, R. K. PAL: Sample-efficient personalized reward modeling for pluralistic alignment. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. The Llama 3 herd of models, 2024.

Ge, T., Chan, X., Wang, X., Yu, D., Mi, H., and Yu, D. Scaling synthetic data creation with 1,000,000,000 personas, 2024.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 6626–6637, 2017.

Hou, Y., Li, J., He, Z., Yan, A., Chen, X., and McAuley, J. Bridging language and items for retrieval and recommendation, 2024.

Hugging Face and Sentence-Transformers. all-MiniLM-L6-v2: Sentence embedding model. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2, 2021. Accessed: 2026.

Kirk, H. R., Whitefield, A., Röttger, P., Bean, A., Margatina, K., Ciro, J., Mosquera, R., Bartolo, M., Williams, A., He, H., Vidgen, B., and Hale, S. A. The PRISM alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2024a.

Kirk, R., Mediratta, I., Nalmpantis, C., Luketina, J., Hambro, E., Grefenstette, E., and Raileanu, R. Understanding the effects of RLHF on LLM generalisation and diversity. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024b.

Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J. E., Zhang, H., and Stoica, I. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP)*, pp. 611–626. Association for Computing Machinery, 2023.

Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., and Aila, T. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Lee, S., Park, S. H., Kim, S., and Seo, M. Aligning to thousands of preferences via system message generalization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

McInnes, L., Healy, J., Saul, N., and Großberger, L. UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018.

Naeem, M. F., Oh, S. J., Uh, Y., Choi, Y., and Yoo, J. Reliable fidelity and diversity metrics for generative models. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pp. 7176–7185, 2020.

Ni, J., Li, J., and McAuley, J. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th*

International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 188–197, Hong Kong, China, 2019. Association for Computational Linguistics.

Padmakumar, V. and He, H. Does writing with language models reduce content diversity? In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*, pp. 1–22, San Francisco, CA, USA, 2023. Association for Computing Machinery.

Reimers, N. and Gurevych, I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China, 2019. Association for Computational Linguistics.

Rendle, S., Freudenthaler, C., Gantner, Z., and Schmidt-Thieme, L. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 452–461, Montreal, Canada, 2009.

Salemi, A., Mysore, S., Bendersky, M., and Zamani, H. LaMP: When large language models meet personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 7370–7392, 2024.

Salewski, L., Alaniz, S., Rio-Torto, I., Schulz, E., and Akata, Z. In-context impersonation reveals large language models' strengths and biases. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., and Hashimoto, T. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pp. 29971–30004, 2023.

Sorensen, T., Moore, J., Fisher, J., Gordon, M. L., Mireshghallah, N., Rytting, C. M., Ye, A., Jiang, L., Lu, X., Dziri, N., Althoff, T., and Choi, Y. Position: A roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al. Qwen3 technical report, 2025.

Zollo, T. P., Siah, A., Ye, N., Li, A., and Namkoong, H. PersonalLLM: Tailoring LLMs to individual preferences. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

A. Hyperparameters and Implementation Details

Table 4 gives the preference-encoder hyperparameters used to embed the roughly 268K real Amazon-Reviews users. Table 5 covers the persona-embedder and SSR settings, and Table 6 covers the synthetic-baseline settings.

Table 4. Encoder hyperparameters used to learn $u_i^* \in \mathbb{R}^d$.

HYPERPARAMETER	VALUE
NUMBER OF PROTOTYPES K	4
OUTPUT DIMENSION d	128
PROJECTOR HIDDEN WIDTH	256
PROJECTOR HIDDEN LAYERS	2
DROPOUT	0.1
SCORE TYPE	COSINE
LOGIT-SCALE INIT	$\log(1/0.07)$
LOSS	CROSS-ENTROPY ON PAIRS
PROJECTOR LR	1×10^{-4}
MIXTURE-WEIGHT LR ($\mathbf{w}$)	5×10^{-3}
LOGIT-SCALE LR	0.1
WEIGHT DECAY	0.01
BATCH SIZE	512
MAX EPOCHS	20
PATIENCE (EARLY STOP)	5
LR SCHEDULER	ONECYCLELR
WARMUP FRACTION	10%
PRECISION	FP16-MIXED
GRADIENT CLIP	1.0
SEED	42

Table 5. SSR and persona-embedder hyperparameters.

HYPERPARAMETER	VALUE
PROBE ITEMS PER PERSONA	100
HEAD/TORSO/TAIL RATIO	20 / 40 / 40
PERSONA-EMBEDDING OPTIMIZER	ADAM
PERSONA-EMBEDDING STEPS	200
PERSONA-EMBEDDING LR	0.01
PERSONA GENERATION T	0.9
PERSONA GENERATION TOP-p	0.95
PERSONA GENERATION CAP	512 TOKENS
SSR RATING T	0.1
SSR RATING TOP-p	0.95
SSR RATING CAP	8 TOKENS
PAIR MINING THRESHOLD	$\lvert r_a - r_b \rvert \geq 1$

B. Cross-split distributional metrics

The full distributional metric panel on the three single-category splits is in Table 7, Table 8, and Table 9. Column structure mirrors Table 2 in the main body; each row is averaged over five resamples of size 500, and the `MMD p` column reports a permutation p-value with $B = 500$ permutations.

C. Cross-split FGI scores

Table 10 gives the per-split, per-method FGI panel, derived from the per-split distributional metrics through the normalization defined in Section 3.4. Higher FGI is farther from a real bootstrap, FGI=1 matches a moment-matched random baseline, and FGI>1 flags worse-than-random behavior on that metric.

Table 6. Synthetic-baseline hyperparameters.

HYPERPARAMETER	VALUE
DIRICHLET CONCENTRATION β	1.0
DIRICHLET ANCHORS n_{anchor}	3
GMM COMPONENTS K_{GMM}	16
GMM COVARIANCE TYPE	FULL
PROTOTYPE PERTURBATION σ	0.01
SUBSAMPLE SIZE PER DRAW	500
RESAMPLES FOR DISTRIBUTIONAL METRICS	5
MMD PERMUTATIONS B	500

Table 7. Musical_Instruments split distributional metrics. n=500 per row.

METHOD	FPD	MMD	MMD P	COV.	SWD	COS MEAN
REAL_SUBSAMPLE	6.0E-06	2.6E-04	0.295	0.978	1.7E-04	0.727
SYNTHETIC_DIRICHLET	4.2E-05	3.19E-03	0.012	0.960	4.13E-04	0.805
SYNTHETIC_GMM	1.33E-04	9.39E-03	0.002	0.156	5.33E-04	0.685
SYNTHETIC_PROTOTYPE	3.12E-02	0.406	0.002	0.000	1.23E-02	0.031
LLM_GROUNDED (LLAMA)	4.43E-04	0.421	0.002	0.084	1.18E-03	0.985
LLM_GROUNDED (QWEN)	5.09E-04	0.462	0.002	0.066	1.21E-03	0.986
RANDOM_BASELINE	1.11E-03	4.86E-02	0.002	0.000	1.01E-03	0.615

D. Persona prompts

The three persona-generation prompts and the SSR rating prompt are reproduced verbatim below. Slot names in braces are filled at runtime from per-persona attribute or seed-history dictionaries.

(a) Free-form Demographic. System prompt:

```
You are tasked with creating a realistic, detailed persona for a consumer.
Generate a persona with the following demographic attributes.  Be specific
and avoid stereotypes.  The persona should feel like a real individual
with nuanced preferences.
```

User prompt template:

```
Age: {age}.  Gender: {gender}.  Location: {location}.  Occupation:
{occupation}.  Income bracket: {income}.  Education: {education}.  Write
a 150 to 250 word persona for this person, including tastes, daily habits,
and shopping priorities.
```

(b) Free-form Backstory. System prompt:

```
You are a creative writer generating a rich, detailed backstory for a
consumer persona.  The backstory should include life history, current
situation, hobbies, values, and specific experiences that shape their
purchasing preferences.  Make this person feel real, complex, and unique.
```

User prompt template:

```
Write a 200 to 300 word backstory for a consumer.  Do not list demographic
attributes; instead, weave them naturally into a narrative.  End with a
short paragraph on what kinds of products they tend to buy and why.
```

Table 8. Software split distributional metrics. n=500 per row.

METHOD	FPD	MMD	MMD P	COV.	SWD	COS MEAN
REAL_SUBSAMPLE	2.12E-05	-6.27E-05	0.431	0.960	3.78E-04	0.658
SYNTHETIC_DIRICHLET	2.89E-04	1.88E-02	0.002	0.884	1.06E-03	0.787
SYNTHETIC_GMM	1.48E-04	4.34E-04	0.255	0.500	5.58E-04	0.662
SYNTHETIC_PROTOTYPE	1.66E-02	0.265	0.002	0.000	8.31E-03	0.155
LLM_GROUNDED (LLAMA)	2.64E-03	0.478	0.002	0.088	3.20E-03	0.977
LLM_GROUNDED (QWEN)	2.64E-03	0.465	0.002	0.078	3.18E-03	0.978
RANDOM_BASELINE	4.73E-03	3.63E-02	0.002	0.000	1.60E-03	0.604

Table 9. Video_Games split distributional metrics. n=500 per row.

METHOD	FPD	MMD	MMD P	COV.	SWD	COS MEAN
REAL_SUBSAMPLE	2.93E-05	-2.40E-04	0.505	0.932	2.92E-04	0.675
SYNTHETIC_DIRICHLET	1.47E-04	1.40E-02	0.002	0.906	6.61E-04	0.767
SYNTHETIC_GMM	1.60E-04	6.85E-04	0.186	0.280	4.90E-04	0.667
SYNTHETIC_PROTOTYPE	2.65E-02	0.371	0.002	0.000	1.09E-02	0.051
LLM_GROUNDED (LLAMA)	1.19E-03	0.368	0.002	0.060	1.80E-03	0.983
LLM_GROUNDED (QWEN)	1.28E-03	0.430	0.002	0.046	1.93E-03	0.988
RANDOM_BASELINE	2.25E-03	3.42E-02	0.002	0.000	1.18E-03	0.563

(c) Behaviorally Grounded. System prompt:

```
You are tasked with creating a realistic consumer persona based on a real
person's purchase history. The following data shows their actual shopping
behavior in {category_label}. Generate a detailed persona that faithfully
reflects this person's tastes, priorities, and decision-making style. Do
NOT invent preferences that contradict the evidence. Stay grounded in
what the data reveals.
```

User prompt template:

```
Top-rated items by this user (rating, category, title, review snippet):
{top_items_block}. Empirical rating distribution: {rating_hist}. Total
reviews: {n_reviews}. Write a 150 to 250 word persona that explains this
purchasing pattern. Mention concrete categories or products the user
clearly favors.
```

SSR rating prompt. System prompt:

```
You are role-playing as the following person: {persona_description}. You
must respond ONLY from this person's perspective. Stay completely in
character.
```

User prompt template:

```
Based on your personal preferences, how likely would you be to purchase
the following product? Title: {item_title}. Category: {item_category}.
Description: {item_desc}. Use a 1 to 5 scale where 1=Definitely would NOT
buy, 2=Probably would not buy, 3=Might or might not buy, 4=Probably would
buy, 5=Definitely would buy. Respond with ONLY a single number.
```

Table 10. Cross-split FGI by method and metric.

SPLIT	METHOD	FGI-FPD	FGI-MMD	FGI-SWD	FGI-COV
COMBINED	REAL_SUBSAMPLE	0.000	0.000	0.000	0.000
COMBINED	SYNTHETIC_DIRICHLET	0.055	0.491	0.484	0.101
COMBINED	SYNTHETIC_GMM	0.012	-0.024	0.034	0.456
COMBINED	SYNTHETIC_PROTOTYPE	2.28	5.97	4.82	1.000
COMBINED	LLM_GROUNDED (LLAMA)	0.437	7.39	1.80	0.911
COMBINED	LLM_GROUNDED (QWEN)	0.463	8.84	1.85	0.905
COMBINED	RANDOM_BASELINE	1.000	1.000	1.000	1.000
MI	REAL_SUBSAMPLE	0.000	0.000	0.000	0.000
MI	SYNTHETIC_DIRICHLET	0.033	0.061	0.289	0.018
MI	SYNTHETIC_GMM	0.115	0.189	0.432	0.840
MI	SYNTHETIC_PROTOTYPE	28.21	8.40	14.44	1.000
MI	LLM_GROUNDED (LLAMA)	0.395	8.71	1.20	0.914
MI	LLM_GROUNDED (QWEN)	0.455	9.55	1.24	0.933
MI	RANDOM_BASELINE	1.000	1.000	1.000	1.000
SOFTWARE	REAL_SUBSAMPLE	0.000	0.000	0.000	0.000
SOFTWARE	SYNTHETIC_DIRICHLET	0.057	0.518	0.556	0.079
SOFTWARE	SYNTHETIC_GMM	0.027	0.014	0.148	0.479
SOFTWARE	SYNTHETIC_PROTOTYPE	3.51	7.30	6.51	1.000
SOFTWARE	LLM_GROUNDED (LLAMA)	0.556	13.16	2.32	0.908
SOFTWARE	LLM_GROUNDED (QWEN)	0.555	12.81	2.30	0.919
SOFTWARE	RANDOM_BASELINE	1.000	1.000	1.000	1.000
VG	REAL_SUBSAMPLE	0.000	0.000	0.000	0.000
VG	SYNTHETIC_DIRICHLET	0.053	0.413	0.417	0.028
VG	SYNTHETIC_GMM	0.059	0.027	0.223	0.700
VG	SYNTHETIC_PROTOTYPE	11.92	10.79	12.01	1.000
VG	LLM_GROUNDED (LLAMA)	0.521	10.69	1.71	0.936
VG	LLM_GROUNDED (QWEN)	0.564	12.50	1.85	0.951
VG	RANDOM_BASELINE	1.000	1.000	1.000	1.000

E. Algorithm: Persona embedding via SSR

Algorithm 1 gives the SSR-then-optimize procedure that places each LLM persona into the same $\mathbb{R}^d$ as the real users. The encoder, the projector family, the logit scale, and the reference context $\bar{\mathbf{c}}$ are all frozen at the values learned during encoder training; only the per-persona mixture weight $\mathbf{w}$ is updated. The procedure is identical for all three persona conditions and uses the hyperparameters in Table 5.

Algorithm 1 Persona embedding via SSR (full procedure).

Input: persona description D, frozen sentence encoder ϕ, frozen projectors $\{P_k\}_{k=1}^{K}$, frozen logit scale e^{σ}, reference context $\bar{\mathbf{c}}$, item pool $\mathcal{I}$, mixture temperature τ
Output: persona embedding $\mathbf{u} \in \mathbb{R}^d$
Step 1. Sample $M = 100$ stratified probe items $\mathcal{P} \subset \mathcal{I}$ at head/torso/tail ratio 20/40/40 with k-means diversification inside each band.
Step 2. For each $p \in \mathcal{P}$, query the LLM at T=0.1 with the SSR rating prompt conditioned on D, recording integer rating $r_p \in \{1, \ldots, 5\}$.
Step 3. Build pair set $\mathcal{B} \leftarrow \{(a, b) : a, b \in \mathcal{P},\ r_a - r_b \geq 1\}$.
Step 4. Initialize $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.
for $t = 1$ **to** 200 **do**
  Compute $u_{\mathbf{w}}(\bar{\mathbf{c}}) = \sum_k \mathrm{softmax}(\mathbf{w}/\tau)_k \, P_k(\bar{\mathbf{c}})$.
  Compute $u_{\mathbf{w}}(\phi(t))$ for each item t touched by $\mathcal{B}$.
  Compute $s_{\mathbf{w}}(a)$, $s_{\mathbf{w}}(b)$ via cosine score and e^{σ} for $(a, b) \in \mathcal{B}$.
  $\mathcal{L}(\mathbf{w}) \leftarrow \sum_{(a,b)\in\mathcal{B}} -\log \mathrm{sig}\big(s_{\mathbf{w}}(a) - s_{\mathbf{w}}(b)\big)$.
  $\mathbf{w} \leftarrow \mathrm{Adam}\big(\mathbf{w}, \nabla_{\mathbf{w}}\mathcal{L}\big)$ with lr $= 0.01$.
end for
Step 5. Return $\mathbf{u} \leftarrow \sum_k \mathrm{softmax}(\mathbf{w}/\tau)_k \, P_k(\bar{\mathbf{c}})$.

F. UMAP visualization

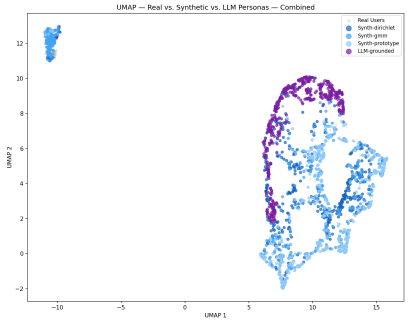

Figure 5. UMAP (McInnes et al., 2018) projection of real users (gray) overlaid with the three synthetic baselines and the grounded LLM personas on the Combined split. Real users form a distinct manifold (gray cluster, top-left); grounded LLM personas (purple arc, top-right) land in a geometrically separate region, while Dirichlet and GMM synthetics span a broader cloud that also fails to overlap the real manifold.

G. Paired distance distributions

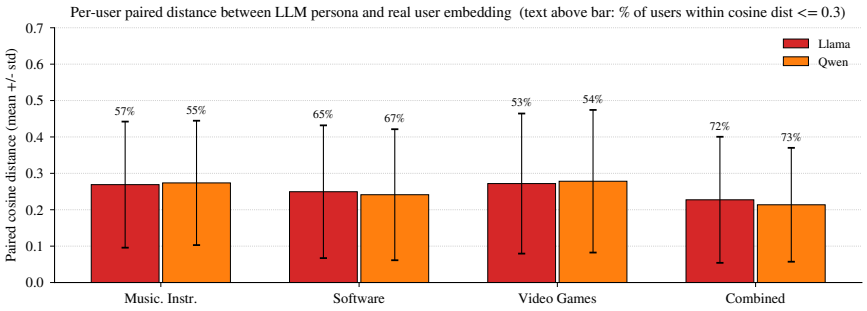

Figure 6. Distribution of per-row paired cosine distance between each grounded persona and its seed user, by split and model. Mass concentrates well below the 0.3 threshold (mean 0.21 to 0.28), so personas land close to their own seed users even though the population coverage in Figure 3 is poor.

H. Norm ratios

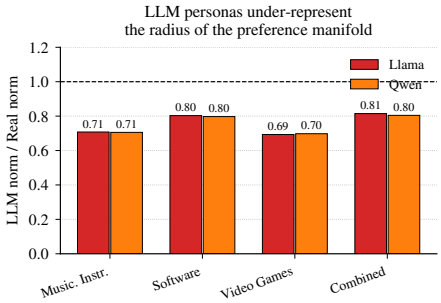

Figure 7. Mean L2 norm of LLM persona embeddings divided by mean L2 norm of real users, per split and model. Ratios of 0.69 to 0.81 indicate persona vectors are systematically shorter than real users; together with high intra-cosine, this places personas inside a tight near-centroid cluster.