BEYOND USER IDENTIFICATION: TARGET-ALIGNED VALIDATION OF PERSONALIZED LLM REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Personalized-LLM evaluations often treat successful user identification, representation separation, or predictive improvement as evidence that a learned user representation captures a named preference. That inference is not valid in general. Users can be distinguishable because of response variability or content exposure even when the target property, such as mean preference direction, is identical. We call this the *target-identification problem*: does the specific user-level property claimed for a representation actually vary across users? This question is distinct from whether users are identifiable or whether personalization is useful.

We introduce *Target-Aligned Validation*, a practical audit that starts from a prespecified user-level property $T(P_u)$. It checks whether the observation design supports estimating that property, compares estimates from separate histories against a content-preserving reference, tests sensitivity to cross-history dependence, and reports finite-history reliability. Behavioural effects and predictive utility are evaluated separately. Controlled experiments show that broad user identification can succeed when every user has the same mean target. Across 133 MultiPref users and three LLMs, the mean-target test detects heterogeneity at all nine tested model-by-history combinations under chronological blocking, conditional on cross-history error assumptions. At the deepest tested history, estimated reliability is only about 5%, and none of the matched utility comparisons establishes a benefit over pooling. In a post-hoc wider-gap analysis, none of the three models passes the prespecified target decision, leaving the role of dependence unresolved. A partial audit of PerFit shows how the procedure transfers to a learned personalization module and which quantities that design cannot support. Target-Aligned Validation separates user identity, target heterogeneity, individual reliability, and downstream utility into distinct evaluation questions.

1 INTRODUCTION

Personalized language models increasingly compress a user's history into embeddings, user-specific parameters, or representation shifts that are then used to adapt future outputs (Liu et al., 2025; Tan et al., 2024; Zhang et al., 2025; Liu et al., 2026). A common evaluation strategy is to ask whether those representations distinguish users, recover their owners, or improve held-out prediction or generation (Qiu et al., 2025b;a; Zheng et al., 2026; Li & Brunskill, 2026). These results establish that a representation contains user-specific or useful information, but they do not by themselves establish that it captures the *particular user property* used to interpret it.

The problem is *target misattribution*: evidence that users are distinguishable can be mistaken for evidence that a named target differs. Consider two users with the same average preference direction but different response variability. Their representation distributions can be easy to distinguish even though the mean target is identical. Likewise, users exposed to different content can appear different because of exposure rather than preference. Without a target-specific test, an evaluation can therefore support the wrong scientific interpretation or motivate an individual estimate that is too noisy to improve on pooling (Shchetkina & Berman, 2024; Li & Brunskill, 2026).

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.

Current work pursues three complementary directions. First, personalization methods learn user embeddings, parameters, or activation shifts and evaluate whether they improve generation (Liu et al., 2025; Tan et al., 2024; Zhang et al., 2025; Liu et al., 2026). Second, methods such as SteerX, NextQuill, DPL, and DEP isolate preference-bearing structure or characterize inter-user differences (Zhao et al., 2025; 2026; Qiu et al., 2025b;a). Third, owner binding, component ablation, and individualized-utility tests ask whether a representation is associated with the right user, causally affects behaviour, or improves decisions (Zhang et al., 2026b; Zheng et al., 2026; Shchetkina & Berman, 2024; Li & Brunskill, 2026). These are useful endpoints, but none is equivalent to asking whether a prespecified functional $T(P_u)$ itself varies across users after accounting for content exposure and finite-history noise.

The missing evaluation layer is target identification. If a representation is interpreted as a mean preference direction, style shift, or another specific property, the evaluation should test that property directly. To our knowledge, there is no general evaluation procedure for personalized representations that starts from a prespecified user-level functional $T(P_u)$, checks whether the observation design supports estimating it, isolates reproducible target variation from content-induced structure, tests sensitivity to cross-history dependence, and separates population heterogeneity from the reliability and utility of an individual estimate. Appendix R maps representative recent evaluations to these distinct questions.

Our approach. Target-Aligned Validation takes three inputs: user histories H_u, a representation extractor R, and a prespecified target $T(P_u)$. It returns a target-specific report containing observation support, an estimated target effect with uncertainty, finite-history reliability, and unresolved assumptions; behavioural effects and predictive utility are reported separately. The audit first checks whether the available observations can distinguish user effects from content exposure, then estimates the same target from separate histories, compares their recurrence with a content-preserving reference, and stress-tests dependence. Figure 1 summarizes the audit architecture, inputs, workflow, and outputs. If the observation design cannot support a target, the output is *not estimable*, rather than evidence of no difference.

Our experiments combine controlled simulations with human-preference data. The main activation-based study evaluates the same 133 MultiPref users with Llama-3.1-8B, Qwen2.5-7B, and Mistral-7B-v0.3; Community Alignment, SFF, and PRISM probe larger cohorts, predictive utility, and sparse content overlap, while PerFit/LaMP-2M tests transfer to a learned personalization module. In controlled worlds, broad user identification detects covariance differences in 60–75% of trials even when all mean targets are identical, whereas the mean-target test rejects in at most 2%. On MultiPref, all nine model-by-history mean-target tests detect heterogeneity under chronological blocking, conditional on the required error assumptions. Yet at 64 total judgments per user, reliability is only about 5%, and no matched comparison establishes an accuracy gain over leave-target-out pooling. A post-hoc wider-gap analysis weakens the signal and none of the three models passes the prespecified target rule. SFF gives the converse pattern: predictive utility improves without detected mean-target heterogeneity.

Our contribution is a target-specific inferential workflow for evaluating personalized representations. Starting from the property a representation is claimed to encode, $\theta_u = T(P_u)$, Target-Aligned Validation links that claim to four requirements: observation support, content-controlled recurrence across separate histories, explicit qualification of cross-history dependence, and finite-history reliability. Behavioural effects and downstream utility remain separate endpoints. The individual statistical ingredients have precedents; the contribution is their organization around a prespecified scientific target and the empirical evidence that user identity, target heterogeneity, individual reliability, and utility can lead to different conclusions on the same personalization problem.

2 From user differences to a specific target

Suppose a model extracts a vector from each preference judgment. A user may have a characteristic distribution of these vectors, P_u, on the content considered in the analysis. Showing that two distributions differ is not the same as showing that a *chosen property* of them differs. We make that property explicit as $T(P_u)$. Two possible null hypotheses are

$$H_0^{\mathrm{id}} : P_1 = \cdots = P_K, \qquad H_0^{T} : T(P_1) = \cdots = T(P_K). \tag{1}$$

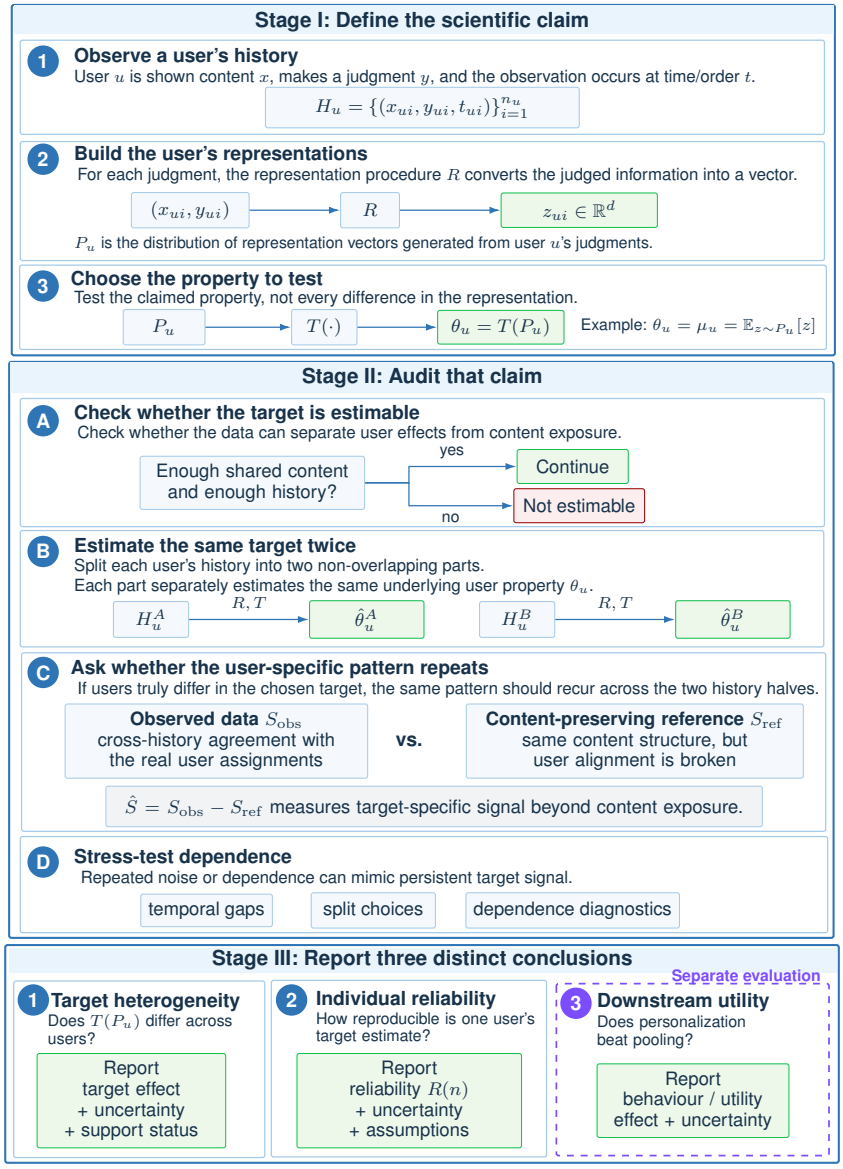

Figure 1: **Target-Aligned Validation as a process.** A user's judgments generate representation vectors and a user-conditioned distribution P_u; the scientific claim specifies a target $\theta_u = T(P_u)$. The audit then checks whether the target is estimable, estimates it from separate histories, compares recurrence with a content-preserving reference, and stress-tests dependence. Population target evidence, individual reliability, and downstream utility answer distinct questions.

Rejecting H_0^{id} tells us that something differs among the distributions, perhaps the average, their variability, or their shape. Rejecting H_0^T is evidence about the specific quantity of interest, provided its estimator and assumptions are valid. Correct owner matching, an effect from intervening on the model, and a predictive gain answer other questions. None is equivalent to testing H_0^T.

For a concrete example, we use activation-based preference steering. A user sees prompt x, chooses response y^+, and rejects the remaining $M-1$ responses. From the residual-stream representation h_ℓ of a frozen model, we compute

$$z = h_\ell(y^+) - \frac{1}{M-1}\sum_{j=1}^{M-1} h_\ell(y_j^-). \tag{2}$$

The same response uses the same cached activation regardless of who judges it. We study two different properties of the resulting user distribution:

$$T_{\text{mean}}(P_u) = \mu_u, \qquad T_{\text{disp}}(P_u) = \mathbb{E}[\|z - \mu_u\|_2^2 \mid u] = \text{tr}(\Sigma_u). \tag{3}$$

We write $\theta_u = T(P_u)$ for the chosen target value and $\hat{\theta}_u$ for an estimate from finite user history. For the mean target used in our main experiments, $\theta_u = \mu_u$ and $\hat{\theta}_u = \hat{v}_u$. The first target above is the average preference direction. The second is the spread around that average. Users can have identical averages but different spreads, or identical averages and spreads but differently oriented covariance. A user-identification test may detect any of these differences, but it cannot say which one produced its result. When we turn to PerFit in Section 8, we define a target appropriate to its learned shift rather than assuming that all personalization methods encode the same object.

3 Target-Aligned Validation

Target-Aligned Validation is a protocol rather than a single universal statistic. It takes three inputs: user histories, a representation extractor, and a prespecified target $T(P_u)$. For each target, the analyst chooses an estimator and reference appropriate to that target and observation design. The protocol then asks whether the target is estimable, whether it recurs across separate histories beyond content-induced structure, how sensitive that evidence is to cross-history dependence, and how reliably it can be estimated from finite history. We instantiate this protocol for mean preference direction in the main experiments, for within-user dispersion in Appendix E, and for PerFit's learned shift in Section 8. Behavioural effects and downstream utility are evaluated separately.

Step 1: Check whether the data support the target. In our implementation, exact-content overlap is used to distinguish response differences from exposure differences without imposing an additional model of content effects. We group identical comparisons into content strata and use a stratum only when at least two users judged it. This is a design choice for the present audit, not a universal requirement of the protocol; another implementation could instead use a separately validated model of content effects. Each included user needs at least $2n$ eligible observations, divided into two histories of n observations each. A dispersion estimate additionally needs $n \geq 2$. We also check that within-content label reassignment is possible. If any required condition fails, the corresponding result is *not estimable under this design*. Passing these conditions does not guarantee statistical power; the eligible number of users and available history still matter. The full operational protocol is in Section I.1.

Step 2: Estimate the specified property twice. For the mean target, divide each user's eligible history into disjoint parts A and B, yielding two estimates of the same target, $\hat{\theta}_u^A = \hat{v}_u^A$ and $\hat{\theta}_u^B = \hat{v}_u^B$. We compare the difference between two users in one half with their difference in the other:

$$D_{uv} = (\hat{v}_u^A - \hat{v}_v^A)^\top (\hat{v}_u^B - \hat{v}_v^B). \tag{4}$$

If the true means differ, the difference should recur across histories. Using the same observations for both sides would also count sampling noise as a user difference. Cross-estimation removes that inflation only under the following assumption.

Assumption 1 (Cross-split error structure)**.** *Write $\hat{v}_u^A = \mu_u + e_u^A$ and $\hat{v}_u^B = \mu_u + e_u^B$. Conditional on the evaluated users and content design, $\mathbb{E}[e_u^A] = \mathbb{E}[e_u^B] = 0$ and $\mathbb{E}[e_u^A (e_v^B)^\top] = 0$ for all evaluated u, v.*

Proposition 1 (Cross-fitted target distance)**.** *Under Assumption 1, $\mathbb{E}[D_{uv}] = \|\mu_u - \mu_v\|_2^2$.*

This is an established cross-validated-distance principle (Walther et al., 2016; Diedrichsen et al., 2016), used here to test a target specified before examining user separation. Appendix B gives the proof.

Step 3: Remove differences attributable to content. Two histories can agree because the same users repeatedly encountered similar material. We construct a reference by reassigning user labels only within identical-content strata. In the notation of Figure 1, S_{obs} is the observed cross-history statistic and S_{ref} is its content-preserving reference. We subtract the latter from the former to obtain

Table 1: Datasets play different roles. Shared-content fractions are measured separately on build and test observations. Cohort definitions are in Appendix A.

Setting	K	Med. hist.	Shared build	Shared test	Purpose
MultiPref (fixed)	133	100	.946	.948	Target, utility, intervention
Community Alignment A	1,929	20	1.000	.000	Larger-cohort target test
SFF (fixed)	54	296	1.000	1.000	Target and utility contrast
Community Alignment B	400	104	.064	.025	Limited-overlap study
PRISM	1,396	19	.008	–	Support check
PerFit / LaMP-2M	94	–	–	.59	Learned-shift audit

the target increment, $\hat{S} = S_{\text{obs}} - S_{\text{ref}}$. This keeps content exposure in the comparison instead of treating it as a preference difference.

A within-content permutation tests equality of *whole conditional distributions*, which is stronger than equality of the chosen mean target. When means agree but dispersions differ, it can therefore reject for the wrong reason if interpreted as a mean test. Our primary mean-target decision instead uses a one-sided user bootstrap on the content-controlled per-user increments, with Monte Carlo uncertainty in the estimated reference included in the decision rule. We resample users, not user pairs. Section I.3 gives the exact rule and calibration experiments.

Step 4: Check dependence rather than assuming it away. If estimation errors persist across histories, repeated noise can look like a stable user target. With $C_{uv} = \mathbb{E}[e_u^A (e_v^B)^\top]$, the expectation becomes

$$\mathbb{E}[D_{uv}] = \|\mu_u - \mu_v\|_2^2 + \operatorname{tr}(C_{uu} - C_{uv} - C_{vu} + C_{vv}). \tag{5}$$

Chronological blocking can reduce short-range dependence, but cannot prove that the covariance terms vanish, especially for persistent user-specific noise. We test sensitivity to splitting and temporal separation and state positive results conditionally on Assumption 1.

Step 5: Report reliability, then evaluate usefulness separately. A population difference can be detectable even when a vector estimated for one user is mostly noise. We report the content-controlled signal $\hat{S}(n)$ as a fraction of the observed between-user variance $\hat{V}(n)$ in the finite-history estimates:

$$R(n) = \frac{\hat{S}(n)}{\hat{V}(n)}. \tag{6}$$

Here n is the number of judgments *in each half*; a total history of $2n$ gives $R(n)$. Under the stated split-history random-effects model, $R(n) = r/(1+r)$ is the fraction of estimated-target variance attributable to persistent signal, the classical reliability form (Shrout & Fleiss, 1979; Spearman, 1910; Brown, 1910). It is not an accuracy gain, a probability of being correct, or evidence of semantic preference by itself. We report target effect and uncertainty, $R(n)$, support status, and unresolved assumptions together. Held-out prediction and interventions then test whether using the representation helps, without standing in for target identification.

4 Experimental setup

We use each dataset to answer a different question (Table 1). MultiPref has shared comparisons and timestamps, allowing us to evaluate target differences, reliability, prediction, and steering on the same 133 users. Community Alignment Track A tests a larger cohort, but without timestamps; Track B and PRISM reveal what happens when exact-content overlap is sparse. SFF provides longer histories and a contrasting utility result. PerFit lets us apply the audit to a learned user module rather than a hand-constructed steering vector.

We extract frozen response-difference activations from Llama-3.1-8B (layer 22), Qwen2.5-7B (layer 20), and Mistral-7B-v0.3 (Jiang et al., 2023) (layer 22). Here n always means observations *per history half*; matched utility uses $n_{\text{total}} = 2n$. Uncertainty uses 10,000 user-bootstrap resamples. User pairs are not resampled independently. Appendix A records model and dataset revisions, eligibility rules, and analysis chronology.

5 User identification does not identify the target

Known targets expose the problem. We first generate data for which we know exactly what varies between users. Figure 2a shows that an identity test responds to changes in mean, dispersion, or

(a) Same identity evidence, different target truths

	Identity	Mean target	Dispersion target
No difference	.050	.050	.033
Mean differs	1.000	1.000	.033
Mean differs (large)	1.000	1.000	.033
Dispersion (2×)	.600	.033	1.000
Dispersion (4×)	1.000	.017	1.000
Cov. orientation	.600	.000	.017
Cov. orientation (strong)	.750	.000	.017

(b) Safeguards solve different failures

	Neither	Cross-fit	Content	Both
null worlds — false-positive rate				
Unequal noise	1.000	.035	.000	.040
Anisotropic cov.	1.000	.015	.025	.020
Exposure only	1.000	1.000	.000	.010
alternative worlds — detection rate				
Mean, R = .03	1.000	.915	.000	.910
Mean, R = .09	1.000	1.000	1.000	1.000

Figure 2: Controlled experiments, $\alpha = .05$. (a) Different data-generating mechanisms can all support user identification even when the mean target is shared. Seven worlds, 60 replicates each, $K = 300$, $n = 8$ per half, $d = 64$. (b) Cross-fitting addresses finite-sample noise, content matching addresses exposure, and their combination addresses both. Five worlds, 200 replicates each, $K = 500$, $n = 8$ per half, $d = 64$. Upper rows are false-positive rates; lower rows are detection rates. Experiment files are identified in the supplement.

covariance orientation. In the two covariance-orientation settings, mean and dispersion are identical across users, but identity tests reject in 60–75% of replicates. The tests of the specified mean and dispersion targets reject in at most 2%. User identification is not incorrect here; interpreting it as evidence of a different average preference would be incorrect.

Why both safeguards are needed. Cross-fitting alone still detects exposure-only differences in all simulated trials. Matching content alone has no detection power for the weaker injected mean difference, whereas combining the two detects it in 91% of trials (Figure 2b). An empirical null built from real activations also illustrates the distinction: broad identity rejects in 6 of 40 trials, while the mean-target test rejects in 0 of 40 ($K = 1{,}929$, $n = 8$). Appendix G evaluates additional representation diagnostics. Their ability to describe user structure does not make them tests of a particular mean.

Dependence remains an essential limitation. In equal-mean simulations, cross-history error correlation of $\rho = .05$ raises the false-positive rate of the mean procedure from .034 at $\rho = 0$ to .980. Chronological blocking is therefore a precaution, not proof of independence. Appendix I separates dependence mechanisms.

6 Heterogeneity, Reliability, and Utility on the Same Users

First question: do their average directions differ? Using the fixed MultiPref cohort and chronological halves, the mean-target procedure detects heterogeneity in all nine model-by-history settings, $n_{\text{total}} \in \{16, 32, 64\}$. These decisions follow the prespecified bootstrap rule and within-model Holm families, and remain *conditional on Assumption 1*. A post-hoc nine-cell Holm sensitivity leaves the decisions unchanged. A separate post-hoc check at three Llama layers also detects the target (Appendix L).

Second question: can we estimate one person's direction reliably? Not at the observed history sizes. At 64 total judgments per user, $R(32)$ is 4.56%, 5.14%, and 4.47% for Llama, Qwen, and Mistral (Figure 3a). Under the split-history model, most of the variation in an individual estimate is finite-history noise. Population-level detection therefore says little about the precision of one user's vector. This low reliability belongs to the mean-direction representation, estimator, history length, and assumptions evaluated here, not to personalized representations in general. PerFit's descriptive cross-half ratio (Section 8) is not a comparable estimate of $R(n)$.

Third question: does choosing that person's vector help? No matched-history accuracy comparison establishes an improvement over a pooled vector that excludes the evaluated user (Figure 3b). At the deepest depth, the upper ends of the 95% intervals are +0.85, +0.55, and +1.27 percentage points. These intervals do not demonstrate zero benefit; they show that a benefit has not been established. In the held-out Llama intervention, pooled steering improves on no steering by +.002761 [+.002078, +.003431] in normalized log-probability margin, whereas owner minus pooled steering

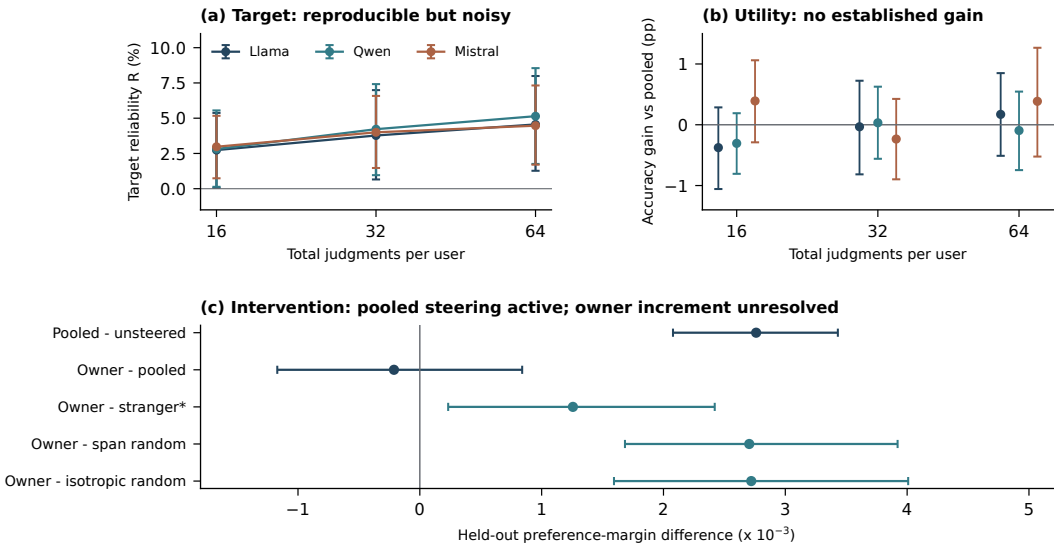

Figure 3: MultiPref, fixed $K = 133$. (a) Mean-target reliability with 95% user-bootstrap intervals; all nine target tests pass their prespecified within-model correction. (b) Held-out accuracy relative to leave-target-out pooling, in percentage points. (c) Llama steering at 64 total judgments, 40 calibration and 93 evaluation users. Owner versus stranger passes the prespecified mean-bootstrap test but not signed-rank sensitivity ($p = .112$); only 54.1% of scored pairs have shared comparisons.

is $-.000212$ $[-.001170, +.000840]$. Owner versus stranger depends on the inference test. Steering can change behaviour without establishing an advantage from choosing a particular user's vector. Together, these observations change how we would report the same experiment: target heterogeneity is detected under specified assumptions, finite-history estimates remain noisy, and an owner-specific utility gain has not been established. None of these statements can replace either of the others.

What happens when we separate histories further? Across 200 random partitions, median $R(32)$ is $(5.6, 6.8, 5.5)\%$ for Llama, Qwen, and Mistral, versus $(4.7, 5.2, 4.5)\%$ with chronological halves. The random-split signal is nearly flat with depth, while chronological signal attenuates, although the split contrasts do not reach conventional significance ($p = .114, .154, .055$). In a *post-hoc* analysis, selecting histories from the earliest and latest thirds increases median separation from 8.00 to 11.06 days. Reliability falls to $(1.95, 2.58, 1.82)\%$ at $n = 16$ per half, and none of the three buffered results passes the prespecified target rule. The selected items change along with the time gap, so this does not isolate dependence or establish preference drift (Section I.4).

What changes across datasets? Community Alignment Track A detects a mean-target difference at $K = 1{,}929$, $n = 8$ per half, with $R = (8.8, 11.5, 8.5)\%$ across the three backbones. Without timestamps, these results depend on random-split independence and are not chronologically blocked confirmation. In Track B, exact matching removes 93.6% of observations as content singletons. Coarse matching still detects broad identity differences in $98.8/95.6/89.6\%$ of matched-K subsamples, but the restricted exact comparison is too weak to interpret non-rejection as identical users. PRISM has similarly limited exact-content support. The dispersion analysis in Appendix E illustrates why a different user-level property requires a different estimator and additional calibration. Its heterogeneity decisions are supported only in the tested simulation settings; its estimated absolute levels and some uncertainty intervals remain design-dependent.

7 Detection and Usefulness Can Diverge

The MultiPref result is not paradoxical: a sufficiently large group can reveal small between-user differences even when individual estimates are noisy. A simple result makes this precise. **Proposition 2 (Population detection can precede reliable individual estimation).** In a Gaussian random-effects model with $\theta_u \sim \mathcal{N}(0, \tau^2)$, $X_{ui} = \theta_u + \epsilon_{ui}$, and $r = n\tau^2/\sigma^2$, population heterogeneity is consistently testable when $r\sqrt{K} \to \infty$. The Bayes-optimal relative reduction in *latent-target squared error* from individualized estimation is $r/(1 + r)$. Thus $K^{-1/2} \ll r \ll 1$ allows detection even

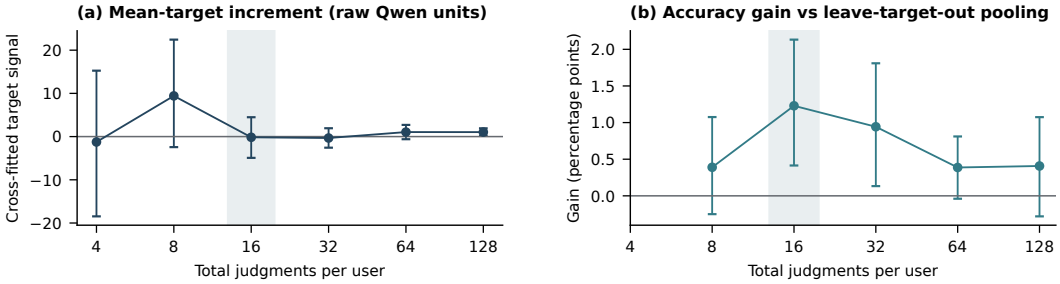

Figure 4: SFF/Qwen, $K = 54$**.** At 16 total judgments, held-out accuracy improves over leave-target-out pooling although the mean-target test does not detect heterogeneity. Error bars are 95% user-bootstrap intervals; the full depth sweep is in Appendix K.

Table 2: PerFit, partial target-aligned audit. Confidence intervals are 95% user-bootstrap intervals. Held-out candidate-tag scores are not directly comparable with PerFit's published free-generation results.

Question	Observation and interpretation
Is fixed-content estimation supported?	94 users, 681 shared descriptions, shared fraction .59.
User-matching evidence	Owner AUC .793; content reference .726; difference $+.068$ $[+.012, +.099]$, $p = .003$. An alternative coupled null does not reject.
Does the fixed-content shift differ?	Increment $+4.38$ $[+1.29, +8.69]$; 70.2% of users have a positive increment.
How reliable is one shift estimate?	**Not estimable** as target reliability. Cross-half ratio .47 [.14, .94] is descriptive only; its numerator and denominator use different content controls.
Does the shift affect held-out scoring?	Owner minus stranger $+.0444$ $[+.0249, +.0667]$; owner minus collective $+.0265$ $[+.0131, +.0409]$.
Can we identify dispersion?	**Not estimable** under the dependence and thin-peer conditions of this design (Appendix Q).

while per-user gains vanish. This does not establish a result about downstream accuracy; proof and controlled phase diagram are in Appendices C and J.

SFF provides the converse empirical pattern (Figure 4). With Qwen and 16 total judgments, held-out accuracy improves by $+1.23$ points $[+.41, +2.13]$ over leave-target-out pooling, with Holm-adjusted $p = .0072$ across six preregistered depths. The mean-target test does not detect heterogeneity ($R = -.1\%$). The slight negative value is a finite-sample estimate and is interpreted as no detected persistent signal, not as literal negative reliability. This is not evidence that true means are equal; the test has limited power at $K = 54$ and this reliability. Improved prediction alone cannot identify the property responsible for the improvement.

8 Can the audit be applied to a learned module?

We apply the procedure to PerFit's learned user-specific shift (Liu et al., 2026). For LaMP-2M, let Q_{shared} be the empirical distribution over 681 shared movie descriptions and define $T(P_u) = \mathbb{E}_{x \sim Q_{\text{shared}}}[\Delta_u(x)]$, the user's mean learned shift on that common content. Shared evaluation observations are 59%. We train shifts from disjoint halves of each of 94 users' histories. This is a *partial* audit; not every step transfers unchanged to a trained module.

The target varies across the evaluated users under the tested design (Table 2), and the held-out scoring contrasts are positive. This is a separate audit of an existing learned method, not a reproduction or correction of its original evaluation conclusions. We cannot, however, separate every persistent user-specific component from the intended semantic component, turn the cross-half ratio into an individual target reliability, or validate dispersion with this constructor. Those are limits of the available observation design, not evidence that PerFit itself fails. The dispersion estimator for the activation testbed uses different statistics and assumptions, and its estimated *heterogeneity* is more reliable than its absolute magnitude (Appendix E).

9 Related work

Personalized language models use retrieval, user-specific parameters, embeddings, activation vectors, and learned shifts (Salemi et al., 2024; Tan et al., 2024; Liu et al., 2025; Zollo et al., 2025; Zhang et al., 2025; Liu et al., 2026). Activation-based approaches build on Activation Addition and Contrastive Activation Addition (Turner et al., 2023; Rimsky et al., 2024). DPL and DEP characterize inter-user differences under related content (Qiu et al., 2025b;a); SteerX and NextQuill seek to separate preference-bearing information from context during construction (Zhao et al., 2025; 2026); and Preference Heads uses component ablation to test a mechanism's behavioural role (Zhang et al., 2026b). These methods address how personalized information is constructed, separated, interpreted, or used. Our question begins after a representation has been chosen: does a particular prespecified property of that representation vary reproducibly across users on a supported content distribution?

Recent work also questions what personalization evaluations establish, but targets adjacent claims. Zheng et al. (2026) separate user-binding evidence from incremental utility by comparing per-user modality weights with shuffled and global alternatives; this tests whether user-specific parameters matter beyond a shared parameter, not whether an arbitrary functional $T(P_u)$ of a representation distribution differs reproducibly across users. Venkatesh & Kurapath (2026) study a different identifiability problem: behaviourally equivalent steering vectors need not be geometrically unique. Theory-grounded style evaluation and actionable-heterogeneity tests likewise address semantic validity or the benefit of individualization rather than the target-identification null used here (Sawant, 2026; Shchetkina & Berman, 2024; Li & Brunskill, 2026).

Our cross-fitted statistic builds on established cross-validated distances (Walther et al., 2016; Diedrichsen et al., 2016), our reliability interpretation is classical, and the distinction between exchangeability and weaker permutation nulls is known (Janssen, 1997; Chung & Romano, 2013). The methodological contribution is the evaluation layer that starts from a researcher-specified $T(P_u)$ and links its interpretation to observation support, content-controlled cross-history recurrence, dependence qualification, and finite-history reliability, while keeping behavioural effects and downstream utility as separate endpoints. Appendix R gives a source-level comparison of representative personalization evaluations; the comparison is descriptive rather than a quantitative reanalysis.

10 Discussion and conclusion

A personalized vector can carry real information and still support the wrong interpretation. Our controlled experiments show why: user identification may respond to differences in variability when average preference directions are the same. MultiPref shows another distinction: a population difference can be detectable while an individual estimate remains too noisy to improve on a shared direction. SFF shows that a prediction gain need not identify which property produced it. Personalization evidence is therefore not interchangeable: knowing *who* a representation belongs to, *what property* differs across users, *how reliably* that property can be estimated, and *whether using it helps* are four different questions.

Target-Aligned Validation makes these distinctions operational. Before interpreting a representation, specify the property of interest and check whether overlapping content supports its estimation. Compare separate histories against a content-preserving reference, report effect and uncertainty, and quantify finite-history reliability. Test behavioural effects and utility with their own endpoints. A result of *not estimable* reveals a data-collection limitation, not an absence of user differences.

The limits are substantive. Correlated errors can inflate cross-history reproducibility, chronological blocking cannot eliminate every kind of shared error, and exact-content matching may discard most observations. In our post-hoc MultiPref buffered-gap check, all three effects attenuate and fail the primary decision rule, but the changed time gap and changed content cannot be disentangled. The mechanism remains unresolved. Community Alignment lacks within-user timestamps, and the PerFit audit cannot establish every target or reliability quantity. Larger models, independent-session fixed-content panels, and methods for sparse-overlap data remain open tests of the framework. The practical recommendation is therefore simple: state the user-level property being claimed, verify that the observations support its estimation, test whether that property recurs beyond content structure, quantify finite-history reliability, and evaluate downstream utility separately.

ETHICS STATEMENT

We use previously released human-preference and personalization datasets without contacting participants. Analyses use pseudonymous user identifiers; we do not attempt to identify individuals or release user representations. We report aggregate results and do not interpret representation differences as personality traits or other sensitive attributes without separate behavioural evidence. We follow the datasets' released terms.

REPRODUCIBILITY STATEMENT

The supplementary code and provenance files specify dataset and model revisions, layers, cohort selection, content strata, null construction, inference settings, and result-generating commands. The SFF fixed-cohort sweep, dependence study, diagnostic stress tests, and $K \times n$ phase diagram were frozen before outcomes were inspected. The three-backbone MultiPref extension was fixed before Qwen and Mistral activations existed; its Llama arm repeats a known same-design result. The MultiPref intervention was frozen before its runner existed. The wider nine-cell Holm analysis, Track-B matched-K control, reliability interpretation, layer checks, random-versus-chronological diagnostics, and buffered-gap sensitivity are post-hoc and identified as such. The earlier scaled Community Alignment intervention plan was fixed after launch but before outcomes were available. PerFit uses a pinned public commit without changing its learning algorithm.

AI USE STATEMENT

Generative AI tools assisted with literature organization, conceptual framing, methodological and experimental-design feedback, code drafting and debugging, result interpretation, figure development, mathematical exposition, and manuscript drafting and editing. The authors reviewed AI-assisted material, verified reported numbers against saved programmatic outputs and provenance records, and are responsible for the final submission and supporting artifacts.

REFERENCES

William Brown. Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 1904–1920*, 3(3):296–322, 1910. doi: 10.1111/j.2044-8295.1910.tb00207.x.

EunYi Chung and Joseph P. Romano. Exact and asymptotically robust permutation tests. *The Annals of Statistics*, 41(2):484–507, 2013. doi: 10.1214/13-AOS1090.

Jörn Diedrichsen, Serge Provost, and Hossein Zareamoghaddam. On the distribution of cross-validated Mahalanobis distances. *arXiv preprint arXiv:1607.01371*, 2016.

Arnold Janssen. Studentized permutation tests for non-i.i.d. hypotheses and the generalized behrens–fisher problem. *Statistics & Probability Letters*, 36(1):9–21, 1997. doi: 10.1016/S0167-7152(97)00043-6.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b, 2023.

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. 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*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/be2e1b68b44f2419e19f6c35a1b8cf35-Abstract-Datasets_and_Benchmarks_Track.html.

Zhaoqi Li and Emma Brunskill. A statistical test for the benefits of personalizing interventions. *Science*, 393(6807):eaeb9506, 2026. doi: 10.1126/science.aeb9506.

Jiahong Liu, Wenhao Yu, Quanyu Dai, Zhongyang Li, Jieming Zhu, Menglin Yang, Tat-Seng Chua, and Irwin King. PerFit: Exploring personalization shifts in representation space of LLMs. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/7a1ac10ea9d98178a389c7fbb3575567-Abstract-Conference.html.

Jiongnan Liu, Yutao Zhu, Shuting Wang, Xiaochi Wei, Erxue Min, Yu Lu, Shuaiqiang Wang, Dawei Yin, and Zhicheng Dou. LLMs + persona-plug = personalized LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9373–9385, 2025. doi: 10.18653/v1/2025.acl-long.461.

Lester James Validad Miranda, Yizhong Wang, Yanai Elazar, Sachin Kumar, Valentina Pyatkin, Faeze Brahman, Noah A. Smith, Hannaneh Hajishirzi, and Pradeep Dasigi. Hybrid preferences: Learning to route instances for human vs. AI feedback. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7162–7200, 2025. doi: 10.18653/v1/2025.acl-long.355. URL https://aclanthology.org/2025.acl-long.355/.

Yilun Qiu, Tianhao Shi, Xiaoyan Zhao, Fengbin Zhu, Yang Zhang, and Fuli Feng. Latent inter-user difference modeling for LLM personalization. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 10599–10617, 2025a. doi: 10.18653/v1/2025.emnlp-main.536.

Yilun Qiu, Xiaoyan Zhao, Yang Zhang, Yimeng Bai, Wenjie Wang, Hong Cheng, Fuli Feng, and Tat-Seng Chua. Measuring what makes you unique: Difference-aware user modeling for enhancing LLM personalization. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 21258–21277, 2025b. doi: 10.18653/v1/2025.findings-acl.1095.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, 2024.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. LaMP: When large language models meet personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7370–7392, 2024. doi: 10.18653/v1/2024.acl-long.399.

Yash Ganpat Sawant. Theory-grounded evaluation exposes the authorship gap in LLM personalization, 2026. URL https://arxiv.org/abs/2604.26460. arXiv:2604.26460.

Anya Shchetkina and Ron Berman. When is heterogeneity actionable for personalization?, 2024. URL https://arxiv.org/abs/2411.16552. arXiv:2411.16552; revised May 2026.

Patrick E. Shrout and Joseph L. Fleiss. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428, 1979. doi: 10.1037/0033-2909.86.2.420.

Charles Spearman. Correlation calculated from faulty data. *British Journal of Psychology, 1904–1920*, 3(3):271–295, 1910. doi: 10.1111/j.2044-8295.1910.tb00206.x.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, 2020.

Zhaoxuan Tan, Qingkai Zeng, Yijun Tian, Zheyuan Liu, Bing Yin, and Meng Jiang. Democratizing large language models via personalized parameter-efficient fine-tuning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6476–6491, 2024. doi: 10.18653/v1/2024.emnlp-main.372.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering, 2023. Introduces Activation Addition (ActAdd); earlier versions circulated as "Activation Addition: Steering Language Models Without Optimization".

Sohan Venkatesh and Ashish Mahendran Kurapath. On the non-identifiability of steering vectors in large language models. *arXiv preprint arXiv:2602.06801*, 2026.

Alexander Walther, Hamed Nili, Naveed Ejaz, Arjen Alink, Nikolaus Kriegeskorte, and Jörn Diedrichsen. Reliability of dissimilarity measures for multi-voxel pattern analysis. *NeuroImage*, 137: 188–200, 2016. doi: 10.1016/j.neuroimage.2015.12.012.

Jinghao Zhang, Yuting Liu, Wenjie Wang, Qiang Liu, Shu Wu, Liang Wang, and Tat-Seng Chua. Personalized text generation with contrastive activation steering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7128–7141, 2025. doi: 10.18653/v1/2025.acl-long.353. URL https://aclanthology.org/2025.acl-long.353/.

Lily Hong Zhang, Smitha Milli, Karen Jusko, Jonathan Smith, Brandon Amos, Wassim Bouaziz, Manon Revel, Jack Kussman, Yasha Sheynin, Lisa Titus, Bhaktipriya Radharapu, Jane Yu, Vidya Sarma, Kris Rose, and Maximilian Nickel. Cultivating pluralism in algorithmic monoculture: The community alignment dataset. In *International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=4NtoAVqfhA.

Weixu Zhang, Ye Yuan, Changjiang Han, Yuxing Tian, Zipeng Sun, Linfeng Du, Jikun Kang, Hong Kang, Xue Liu, and Haolun Wu. Preference heads in large language models: A mechanistic framework for interpretable personalization. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 47742–47754, 2026b. doi: 10.18653/v1/2026.acl-long.2205. URL https://aclanthology.org/2026.acl-long.2205/.

Xiaoyan Zhao, Ming Yan, Yilun Qiu, Haoting Ni, Yang Zhang, Fuli Feng, Hong Cheng, and Tat-Seng Chua. SteerX: Disentangled steering for LLM personalization. *arXiv preprint arXiv:2510.22256*, 2025.

Xiaoyan Zhao, Juntao You, Yang Zhang, Wenjie Wang, Hong Cheng, Fuli Feng, See-Kiong Ng, and Tat-Seng Chua. NextQuill: Causal preference modeling for enhancing LLM personalization. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=xYpVlKMFqv.

Jingyuan Zheng, Xin Zhang, Yang Gu, Dongjing Wang, Yuxiang Wang, Xudong Shen, Haiping Zhang, Youhuizi Li, and Dongjin Yu. Is personalized modality weighting actually personalized? a controlled audit of per-user weighting claims in multimodal recommenders, 2026. URL https://arxiv.org/abs/2608.05655. arXiv:2608.05655.

Thomas P. Zollo, Andrew Wei Tung Siah, Naimeng Ye, Ang Li, and Hongseok Namkoong. PersonalLLM: Tailoring LLMs to individual preferences. In *International Conference on Learning Representations*, 2025.

How to use this appendix. Appendix A records the data and analysis decisions needed to reproduce the audit. Appendices B–E give the statistical arguments and target-specific estimator checks. Appendices F–I examine observation design, diagnostics, and dependence. Appendices J–Q provide additional empirical results, utility comparisons, interventions, and the PerFit transfer. Appendix R maps nearby evaluations to the specific questions they address. The main-text interpretation does not require every technical result to be read in sequence.

A Data, models, and protocol

Datasets and content strata. Community Alignment (Zhang et al., 2026a) is used in two settings. Track A defines content by the exact prompt and candidate-response set, so two observations share a stratum only when annotators faced the same comparison. Track B uses the coarser released grouping (language $\times$ turn index $\times$ pregenerated flag) and is treated as a confounding diagnostic rather than target-specific evidence when exact overlap is sparse. MultiPref (Miranda et al., 2025) and SFF (Stiennon et al., 2020) are pairwise datasets, so Equation (2) has $M = 2$. PRISM (Kirk et al., 2024) appears only in the support audit. Processed-pool fingerprints and source revisions are recorded with the code.

Models and representation extraction. The evaluated backbones are Llama-3.1-8B-Instruct (revision `d10aef79`) at layer 22, Qwen2.5-7B-Instruct (revision `a09a3545`) at layer 20, and Mistral-7B-Instruct-v0.3 (revision `c170c708`) at layer 22. The Llama layer is the frozen primary layer; the other two follow the same frozen 0.70 relative-depth rule. Activations are read from the residual stream at the final token of each response, computed once per prompt-response pair, and reused across annotators.

Cohorts and the meaning of n. In target-identification analyses, n is the number of observations *per split half*; $n = 8$ therefore uses 16 judgments per user. On Community Alignment Track A, the mean-target cohorts contain $K = 2577, 2359, 1929$ users at $n = 2, 4, 8$. Dispersion additionally requires a leave-one-user-out content reference; this removes one user at $n = 2$ and two at $n = 8$, giving $K = 2576$ and $K = 1927$. Utility uses different eligibility rules and a single build estimator: the primary Track-A utility cohort contains $K = 567$ users with at least eight build and eight held-out judgments. The scalar-shrinkage evaluation splits eligible users into calibration and evaluation subsets, yielding $K = 284$ in the primary cell; the hierarchical estimator has its own leakage-free training/calibration/evaluation split and $K = 486$ in that cell. SFF contains 68 eligible utility users, with validation-selected shrinkage scored on an evaluation half of 34 users after tuning on the other half.

Shared content differs by population and split. The support values in Table 1 are deliberately split-specific. Community Alignment Track A has rate 1.000 on the build data used for target identification but 0.000 on its held-out utility items; Track B has $0.064/0.025$. SFF has 1.000 on both. For MultiPref, the full 224-user eligible pool has dataset-level shared-content fractions $0.9694/0.9703$ on train/test, whereas restricting to the fixed $K = 133$ headline cohort removes sharing partners and yields $0.9456/0.9479$. These are the same metric on different populations, not conflicting estimates. The intervention uses another denominator: among its capped evaluation pairs, 0.5406 lie on comparisons also scored by another evaluation recipient, so we describe that endpoint as partially shared-content rather than fully content-controlled.

Finite-history target reliability. For cross-model comparison, let $A_u = \hat{v}_u^A - \bar{v}^A$ and $B_u = \hat{v}_u^B - \bar{v}^B$, and let g_u denote the per-user cross-fitted increment after subtracting its content-preserving reference. We report

$$R(n) = \frac{K^{-1} \sum_u g_u}{\frac{1}{2} \left(K^{-1} \sum_u \|A_u\|_2^2 + K^{-1} \sum_u \|B_u\|_2^2 \right)}. \tag{7}$$

Under the split-history random-effects decomposition in Section D, the numerator estimates persistent between-user target signal and the denominator estimates signal plus finite-history noise; the common $(1 - 1/K)$ centring factor cancels. Thus $R(n) = r/(1 + r)$ and admits a target-estimation reliability interpretation: the fraction of finite-history target-estimate variance attributable to persistent target signal under that model. It remains invariant to global rescaling and orthogonal rotation, which makes it useful for cross-representation comparison at a fixed history depth. The interpretation is conditional on the split-history assumptions and is not downstream utility, a probability of correctness, or a causal effect.

Inference and decision chronology. Primary intervals use 10,000 percentile bootstrap resamples with the user as the resampling unit; pairwise user comparisons are never treated as independent. The original cohort rules, layer choices, depth grids, utility endpoints, and shrinkage grids were fixed before their corresponding headline analyses unless explicitly stated otherwise. Experiment I (the three-model MultiPref replication) and Experiment J (the MultiPref intervention) were frozen before Qwen/Mistral activations or the intervention runner existed, respectively. The preregistered MultiPref multiplicity family is five depths within each model and panel, never across models. The implementation applies Holm to one-sided bootstrap p-values whereas the plan prose names permutation p-values; both families are reported and no decision differs. A wider nine-cell Holm correction across the three models and three headline depths was added *post hoc*; all nine conclusions remain significant (maximum adjusted $p = .0139$ using permutation p-values). The privileged-history arm remains explicitly post hoc.

B Proof of Proposition 1

Let $\delta_{uv} = \mu_u - \mu_v$. Under Assumption 1,

$$D_{uv} = (\delta_{uv} + e_u^A - e_v^A)^\top (\delta_{uv} + e_u^B - e_v^B) \tag{8}$$

$$= \|\delta_{uv}\|_2^2 + \delta_{uv}^\top (e_u^B - e_v^B) + (e_u^A - e_v^A)^\top \delta_{uv} + (e_u^A - e_v^A)^\top (e_u^B - e_v^B). \tag{9}$$

The two linear terms have zero expectation by unbiasedness. The final term expands into

$$\mathbb{E}[(e_u^A)^\top e_u^B] - \mathbb{E}[(e_u^A)^\top e_v^B] - \mathbb{E}[(e_v^A)^\top e_u^B] + \mathbb{E}[(e_v^A)^\top e_v^B], \tag{10}$$

which is zero because each term is the trace of a cross-split error covariance assumed to vanish. Hence $\mathbb{E}[D_{uv}] = \|\mu_u - \mu_v\|_2^2$. Averaging D_{uv} over ordered pairs reduces to a user-level mean, which motivates the user bootstrap rather than treating the $K(K-1)$ pairs as independent.

C Proof of Proposition 2

Let $\bar{X}_u = n^{-1} \sum_i X_{ui}$. Marginally,

$$\bar{X}_u \sim \mathcal{N}\left(0, \tau^2 + \frac{\sigma^2}{n}\right) = \mathcal{N}\left(0, \frac{\sigma^2}{n}(1+r)\right), \qquad r = \frac{n\tau^2}{\sigma^2}. \tag{11}$$

Assume σ^2 is known; this only makes the population test more favorable and keeps the scaling transparent. Under $H_0 : \tau^2 = 0$,

$$Q_K = \frac{n}{\sigma^2} \sum_{u=1}^{K} \bar{X}_u^2 \sim \chi_K^2. \tag{12}$$

A level-α test therefore rejects when $Q_K > q_{K,1-\alpha}$, the $(1-\alpha)$ quantile of χ_K^2. Under the alternative, $Q_K/(1+r) \sim \chi_K^2$. Since $q_{K,1-\alpha} = K + O(\sqrt{K})$ and a χ_K^2 variate is $K + O_p(\sqrt{K})$, the alternative shifts Q_K above the null scale by Kr. Thus the rejection probability tends to one whenever $r\sqrt{K} \to \infty$.

For one user, the Bayes posterior mean is

$$\hat{\theta}_u^{\text{Bayes}} = \frac{\tau^2}{\tau^2 + \sigma^2/n} \bar{X}_u = \frac{r}{1+r} \bar{X}_u. \tag{13}$$

Its Bayes risk is the posterior variance averaged over the data,

$$\mathbb{E}(\hat{\theta}_u^{\text{Bayes}} - \theta_u)^2 = \frac{\tau^2 \sigma^2}{\sigma^2 + n\tau^2} = \frac{\tau^2}{1+r}. \tag{14}$$

If the population mean is known to be zero, the oracle pooled estimator $\hat{\theta}_u^{\text{pool}} = 0$ has risk τ^2. The relative risk reduction from individualization is therefore

$$1 - \frac{\mathcal{R}_{\text{Bayes}}}{\mathcal{R}_{\text{pool}}} = \frac{r}{1+r}. \tag{15}$$

For any sequence satisfying $K^{-1/2} \ll r \ll 1$, the population test is consistent while this relative individualized gain tends to zero. For a fixed-dimensional isotropic vector target, the same argument applies coordinate-wise (or to the summed squared norm), changing constants but not the separation between population evidence and per-user signal-to-noise.

D Reliability interpretation and history scale

The reliability interpretation of $R(n)$ follows from the same split-history random-effects model used in Proposition 2. Let the between-user target covariance have trace τ^2 and let the average per-observation noise covariance have trace σ^2. For an n-judgment target estimate, the expected

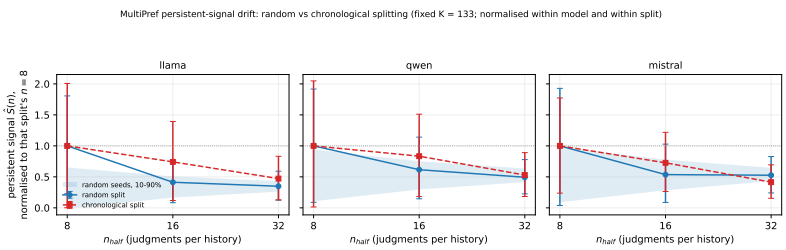

Figure 5: Persistent target signal is stable across repeated random partitions but attenuates under chronological separation. Each panel normalizes signal to the shallow-history level within model. Curves summarize 200 random partitions versus the fixed chronological construction. The same-direction chronological attenuation is suggestive rather than individually significant; identification remains positive in all 18 random/chronological model-by-depth cells.

between-user variance is proportional to $\tau^2 + \sigma^2/n$, while the content-controlled cross-half increment estimates the persistent term τ^2. The shared $(1 - 1/K)$ centring factor cancels, giving

$$R(n) = \frac{\tau^2}{\tau^2 + \sigma^2/n} = \frac{r}{1+r}, \qquad r = \frac{n\tau^2}{\sigma^2}. \tag{16}$$

This equality is a target-estimation statement under the model; it does not convert R into downstream accuracy or causal utility.

We validated the estimator on the repository's known-truth generator before interpreting human results. Estimated R has relative bias within $\pm 1.7\%$ in 11 of 12 tested cells; the worst case is -11.5% in the lowest-SNR shallow cell, and finite-history noise is recovered to within 0.3%. Recomputing the 18 MultiPref human cells reproduces the preregistered Experiment-I values exactly.

The human depth sweep supports the expected $1/n$ noise law: $\widehat{\text{noise}}(n)\, n$ varies by only 4–11% across a fourfold change in n. The shallow $n = 8$ partition had an unusually high signal: its $n = 8$ signal lay at the 100th, 92nd, and 90.5th percentiles of 200 random partitions for Llama, Qwen, and Mistral. Across all 200 partitions, random-split signal is essentially flat, with $\hat{S}(32)/\hat{S}(8) = 1.07, 1.04, 1.07$. By contrast, the chronological estimate declines monotonically to $0.47, 0.53, 0.41$ of its shallow value. The chronological changes fall in the same lower-tail direction for all three representations, but none is conventionally significant against the 200-partition random reference (one-sided $p = .114, .154, .055$), so the result is suggestive rather than confirmatory. Mean temporal separation between the two chronological histories remains approximately eight days at every depth over a 28.9-day corpus span, and per-user time-gap/signal correlations are null. A simple explanation based only on greater temporal separation therefore does not account for the attenuation. Because history depth also changes which observations enter the estimates, the mechanism remains unresolved.

We therefore do not issue a precise sample-size recommendation. Under the optimistic constant-signal extrapolation calibrated at the deepest chronological cell, 50% reliability would require approximately 670 judgments for a Llama target estimate, 590 for Qwen, and 684 for Mistral; the uncertainty ranges are broad and all values are 18–21$\times$ beyond the observed maximum of 32 judgments per estimate. A two-half audit at that reliability would require roughly twice that total history. The random arm better supports the stationary-signal approximation than the chronological arm, but the disagreement between split rules and the extreme extrapolation make these numbers an indicative scale rather than a prescription.

E Dispersion estimator and validation

A mean direction is only one possible scientific target. We include within-user dispersion as a second target to test whether the audit can distinguish variation in average direction from variation in the spread of responses. Dispersion therefore needs its own estimator and calibration, not a new label for the mean test. The heterogeneity decisions below are calibrated in the simulated worlds we tested, whereas absolute dispersion levels and some intervals remain sensitive to content-dependent peer composition.

For observation i of user u in content stratum s_i, let k_{s_i} be the number of users in that stratum and $c_{s_i} = k_{s_i} - 1$. We residualize against the other users in the same stratum,

$$e_{ui} = z_{ui} - \frac{1}{c_{s_i}} \sum_{v \in s_i,\, v \neq u} z_{vs_i}. \tag{17}$$

For two observations i, j of user u in distinct strata, the pairwise squared distance avoids centring on a noisy estimate of μ_u. Under an unbiased content reference and the cross-observation independence condition used by the U-statistic,

$$\frac{1}{2}\mathbb{E}\|e_{ui} - e_{uj}\|_2^2 = \operatorname{tr}\Sigma_u + \frac{1}{2}\operatorname{tr}\bar{T}\left(\frac{1}{c_{s_i}} + \frac{1}{c_{s_j}}\right), \tag{18}$$

where $\operatorname{tr}\bar{T}$ is the within-stratum variance of a single observation around its stratum mean. Estimating and subtracting the second term yields $\hat{q}_u$ for $q_u = \operatorname{tr}\Sigma_u$. Singleton strata are dropped because $c_s = 0$. The pairwise sum is computable in linear time in the number of observations once residuals are available using

$$\sum_{i<j} \|e_i - e_j\|_2^2 = n \sum_i \|e_i\|_2^2 - \left\| \sum_i e_i \right\|_2^2. \tag{19}$$

Table 3: Within-user dispersion on Community Alignment Track A. Cohorts exclude singleton content strata. The calibrated column reports the prespecified heterogeneity decision; increments and their two-sided 95% intervals are not calibrated under non-exchangeable peer composition, and relative SD is descriptive and design-dependent. Effect units differ by model.

Model	n	K	**Decision** *(calibrated)*	Increment [two-sided 95% CI] *(not calibrated)*	rel. SD *(design-dep.)*
Llama-3.1-8B	2	2,576	not identified	$+4.37$ $[-1.60, +10.79]$	.18
Llama-3.1-8B	4	2,359	**identified**	$+4.59$ $[+0.99, +8.40]$	.19
Llama-3.1-8B	8	1,927	**identified**	$+3.54$ $[+0.74, +6.51]$	.16
Qwen2.5-7B	2	2,576	not identified	$+1739$ $[-1237, +4894]$	.16
Qwen2.5-7B	4	2,359	**identified**	$+1997$ $[+214, +3872]$	.17
Qwen2.5-7B	8	1,927	**identified**	$+1419$ $[+100, +2778]$	.14
Mistral-7B-v0.3	2	2,576	not identified	$+0.44$ $[-0.30, +1.23]$	.19
Mistral-7B-v0.3	4	2,359	**identified**	$+0.55$ $[+0.09, +1.03]$	.21
Mistral-7B-v0.3	8	1,927	not identified	$+0.36$ $[-0.0004, +0.73]$	.17

How to interpret the table. The one-sided heterogeneity decision maintained a false-positive rate at or below .024 against nominal .05 in the tested adversarial nulls, with power 1.00 at a dispersion ratio of 1.5; this does not guarantee calibration against other violations. Under non-exchangeable peer composition, the point increment is biased downward and its two-sided interval under-covers (coverage .888–.942 rather than .95), so neither its size nor that interval supports calibrated inference. The absolute level and relative SD are design-dependent. The Mistral $n = 8$ interval's lower endpoint is $-.0004$, while its Monte-Carlo-adjusted decision bound is $-.0083$; neither indicates homogeneity. The Qwen rows used the same depth grid $(2, 4, 8)$, 500 permutations and 10,000 bootstrap samples, at layer 20. We found no positive evidence for the particular assortative-exposure mechanism in the cohorts tested, but the diagnostic cannot exclude other exposure-dependent biases. The derivation and correction comparisons below document these limits rather than treating the point estimates as validated magnitudes.

Across nine content-preserving empirical-null cells, dispersion rejection never exceeds 0.033 at nominal $\alpha = 0.05$. Synthetic injection experiments recover the known post-injection target to within approximately 8%, 4%, and 0.1% across the three tested levels, with interval coverage in all cells. All three real-residual injections are detected, although the magnitude is overestimated by roughly 45–50% at the two largest injections. This is why the main text treats existence of dispersion heterogeneity as more secure than its absolute magnitude.

Two estimator details should be stated precisely before the bias discussion. First, Eq. (18) restricts the pairwise average to observations in *distinct* strata, whereas the $O(nd)$ closed form in Eq. (19) averages over all within-half pairs; the two coincide only when no user contributes two observations to one stratum. We measured that directly rather than assuming it: the share of within-user pairs

that share a stratum is 0 on MultiPref and SFF and 4.6×10^{-5} on Community Alignment Track A, so the shortcut is exact on MultiPref and SFF and immaterial here. It is not safe on designs with repeated within-stratum judgments, and the released code now reports this fraction. Second, the residualisation in Eq. (17) is written against the other *users* of a stratum but implemented against the other *observations*, which again coincide under the same condition.

The asymmetry has a structural source. The exposure correction in the dispersion estimator folds the peer-mean term into $\hat{T}$ by treating stratum co-members as an exchangeable draw. When that fails, exposure is assortative *and* exposure groups differ in mean, the correction under-corrects and $\hat{q}_u$ is biased upward: in controlled worlds where every user has identical true dispersion the bias is $+0.1\%$ under exchangeable co-membership but $+1.8\%$ to $+3.9\%$ under assortative exposure, and it dominates entirely ($+687\%$ to $+1529\%$) when true within-user dispersion is small. The cohort-level heterogeneity *test* is not affected: because the permutation offset preserves each user's stratum profile, the same bias appears in the reference and is subtracted, leaving the false-positive rate at $\leq .024$ against nominal .05 in every adversarial world we constructed, with power 1.00 at a dispersion ratio of 1.5. What does degrade, in two distinct ways, is the point increment and its two-sided interval. The point increment acquires a *downward* bias. The interval *under-covers*, empirical coverage .888–.942 against nominal .95, which means it is too narrow, and therefore anti-conservative rather than conservative. The two failures must be interpreted separately. We therefore separate three things that a single table row would otherwise conflate: the *decision* whether dispersion is heterogeneous, which held its level across the adversarial nulls we simulated and which we report; the *point increment and its two-sided interval*, which are measured but not calibrated in this regime and which we report as such; and the *absolute level* of $\hat{q}_u$, together with relative SD and dispersion reliability, which are design-dependent and whose magnitudes we do not interpret. We also tested the precondition rather than assuming it: an assortativity index (stratum-mean spread of the content-controlled user means, normalised by its exchangeable expectation) is 0.85 on MultiPref, 0.39 on SFF and 0.53 on Community Alignment Track A, in every case *below* a random-relabelling reference of $0.98 \pm .03$, $0.83 \pm .13$ and $0.99 \pm .04$. We therefore find no evidence of excess assortativity in these cohorts. Two limits on what that licenses: the index is computed from leave-one-out residuals, which biases it downward and so makes it conservative for *detecting* assortativity; and it tests one mechanism, assortative exposure with group-differing means, among the several that could make stratum co-membership non-exchangeable. A null here is evidence against that mechanism, not a demonstration that no exposure-dependent bias affects these levels.

E.1 Two attempted corrections, and why neither is adopted

Because the bias above is real, we tried to remove it rather than only bound it. The two approaches address different estimation errors and carry different support and power costs.

Attempt 1: subtract the permutation offset from the level (rejected). If the within-stratum permutation preserved the peer-mean term $(1/2)\|\bar{\delta}_{s_i} - \bar{\delta}_{s_j}\|^2$, then $\tilde{q}_u = \hat{q}_u - \mathbb{E}_\pi[\hat{q}_u] + \hat{T}$ would cancel it with no exchangeability assumption. It does not. Under the permutation the pseudo-owner is drawn from the stratum itself, so $\mathbb{E}[\delta_v] = \bar{\delta}_s$ and the cross-term cancels the squared difference exactly; the permutation *removes* the term rather than carrying it. Measured directly, $\hat{q} - \mathbb{E}_\pi[\hat{q}]$ is -4.016 in an exchangeable world and -4.019 in an adversarial one, indistinguishable, so the offset carries no information about the bias, and the "corrected" value is further from the truth than the original ($+0.146$ and $+0.373$ against -0.001 and $+0.269$). Attempt 1 is invalid and we record it only to close it off.

Attempt 2: a cross-user stratum-pair plug-in (valid, but not adopted). The peer-mean term is identified only where other users occupy the *same pair* of strata. For each within-half pair of user u in strata (s_i, s_j), let $D_{-u}(s_i, s_j)$ be $\|e_i^w - e_j^w\|^2/2$ averaged over users $w \neq u$ present in both. Then $\|e_i^u - e_j^u\|^2/2 - D_{-u}(s_i, s_j) \to \operatorname{tr}\Sigma_u - \operatorname{tr}\bar{\Sigma}$, removing the peer-mean term with no exchangeability assumption. This works as designed: over 300 replicates per world the primary estimator's relative bias of $+690\%$ in the adversarial near-zero-dispersion world falls to -0.004%, while the heterogeneity decision keeps a false-positive rate of .020–.047 against nominal .05 and power 1.000 at dispersion ratios 1.5, 2 and 3.

We do not adopt it because the experiments identify two costs. First, it costs support: the pair must be co-occupied, and on Community Alignment Track A at $n = 2$ the usable cohort falls from $K = 2{,}576$ to $K = 1{,}547$, with 76.3% of pairs retained. Second, and decisively, it centres each user against *their own donor pool* rather than against the population. When dispersion is confounded with exposure , users who see similar content also differ in scatter, the donors are drawn from the same exposure neighbourhood, and the estimator subtracts the very heterogeneity it should detect. In a world where one exposure group has twice the within-user scatter of the other, the primary procedure detects this in 100% of replicates and the plug-in in 8.3%, against a common false-positive rate near nominal. Trading a bias that our exposure diagnostics say is not triggered on these cohorts for a near-total loss of power against a plausible alternative is not a favourable exchange.

What we do instead, and how the two estimators compare. We retain the primary procedure and run the plug-in as a post-hoc robustness check on all six Community Alignment Track A cells. Before comparing them, three things have to be said, because the two are not interchangeable.

They do not estimate the same thing. The primary statistic targets $(1 - 1/K)\,\mathrm{Var}_u(q_u)$ with $q_u = \mathrm{tr}\,\Sigma_u$ centred against the whole cohort, over every eligible user and every within-half pair. The plug-in targets the variance of $\mathrm{tr}\,\Sigma_u - \mathrm{tr}\,\bar{\Sigma}_{\mathrm{donors}(u)}$, dispersion centred against each user's own donor pool, over the users who have donors and the pairs that are co-occupied. At $n = 8$ no user is dropped, but only 74.7% of pairs are retained; at $n = 2$ the cohort itself falls from $K = 2{,}576$ to $1{,}547$. The nulls differ too: the primary decision references a within-stratum identity permutation, i.e. exchangeability given content, whereas the plug-in has no permutation reference at all.

A Monte-Carlo adjustment is not applicable to the plug-in, and that is not a licence to treat the two decisions as equivalent. The $1.645\,\widehat{\mathrm{se}}_{\mathrm{MC}}$ term in our primary rule exists to cover the Monte-Carlo error in an *estimated permutation offset*, which enters every g_u identically and is therefore invisible to a user-level bootstrap; it shrinks as the permutation count grows. The plug-in subtracts no such offset, its reference $D_{-u}(s_i, s_j)$ is a deterministic average over the donors actually present, not a sampled quantity, so the term it would correct is zero by construction. What the plug-in *does* carry is finite-donor sampling error in D_{-u}, shared across users through a common donor pool and so also only partly visible to a user bootstrap. We have not quantified that, and we therefore do not present the plug-in as having a validated decision rule.

The like-for-like comparison. Under the one rule both support, an ordinary two-sided 95% user-bootstrap interval with lower bound above zero, five of the six cells agree. Both fail at $n = 2$ for Llama ($[-1.60, +10.79]$ versus $[-2.17, +16.59]$) and Mistral ($[-0.30, +1.23]$ versus $[-0.44, +1.59]$); both pass at $n = 4$ and, for Llama, at $n = 8$. The single disagreement is Mistral at $n = 8$: primary procedure $[-0.0004, +0.7326]$, plug-in $+0.5407\,[+0.3484, +0.7322]$. Note that -0.0004 is the *ordinary* interval's lower bound. Under our **prespecified primary rule** the same cell's Monte-Carlo-adjusted bound is -0.0083, so the margin of failure under the rule that actually governs the decision is about twenty times larger than the ordinary interval suggests. It is still a failure by a small margin, and we do not read it as evidence of homogeneity; but "fails by four ten-thousandths" would describe the wrong quantity.

Multiplicity. These six cells are a post-hoc descriptive comparison, not a decision family, and we do not fold them into the preregistered within-model families. For transparency: Holm across the six plug-in p-values would leave four significant ($p_{\mathrm{Holm}} = .0006$ for Llama $n \in \{4, 8\}$ and Mistral $n \in \{4, 8\}$), including the disputed cell. We report that rather than adopt it, because choosing a new family after seeing which cells moved is exactly the freedom the preregistration exists to remove.

The two estimators have different limitations, and neither dominates. It would be wrong to present the plug-in's power loss as proof that the primary estimator is better. Each has a demonstrated failure mode and neither dominates. The primary absolute estimator has a known exposure-dependent upward bias; for the *heterogeneity decision* that bias is largely absorbed by the permutation offset, which is why its false-positive rate stayed at $\leq .024$ in the adversarial worlds we built, but "the worlds we built" is the limit of that evidence. The plug-in removes the bias at its source and needs no exchangeability assumption, but centres locally and so cannot see dispersion heterogeneity aligned with exposure, detecting it in 8.3% of replicates where the primary procedure detects it in 100%. Our reason for retaining this estimator as primary is specific to these data rather than general: the exposure diagnostic we ran on these cohorts (assortativity index below its exchangeable reference in

all three) finds no evidence of the mechanism that triggers its bias, though it tests one mechanism and is conservative for detecting even that one; whereas exposure-aligned dispersion heterogeneity is likewise not ruled out and is precisely what the plug-in would miss. We are choosing between two procedures with different unverified exposures, not between a safe one and a risky one. On a corpus with strongly assortative exposure the preference could reasonably reverse. We retain both estimates, and the disagreement, rather than resolving it by assertion.

What this leaves open. The absolute dispersion level remains a design-dependent quantity under non-exchangeable co-membership. We have not found an estimator that removes that dependence at acceptable cost, and the question remains open.

F Proposed protocols for future data collection

Nothing in this section was run for this paper. These are design recommendations that follow from our negative and ambiguous results, stated so that the dependence and support limits we document are actionable rather than merely acknowledged.

Separating persistent user effects from a mean target. Our blocked simulations show that chronological blocking repairs short-range dependence at moderate strength and does nothing against a user-specific offset that persists across the whole observation period (Section I), and the buffered-gap arm, which widens temporal separation, fails our primary rule in all three backbones. Widening the gap within a single collection period cannot resolve this, because the confound is a property of the user, not of the interval. The design that would separate them is a *repeated fixed-content probe*: a small panel of identical items, re-presented to the same users in sessions separated by an interval long enough that session-level state is plausibly independent (weeks rather than minutes), with the probe items held fixed across sessions and interleaved among unrelated items so that they are not recognisably a retest. That yields two estimates of the same user-level functional on *identical* content from independent sessions, which is what Assumption 1 actually requires and what no split of a single session can supply. A drift estimate falls out of the same design as the session-to-session change on fixed content, which our buffered analysis cannot isolate because it changes content and separation together. We propose this; we have not run it, and we do not infer what it would show.

Sparse-content corpora. The support gate returns "not estimable" on designs such as PRISM (0.008 shared build content) and Community Alignment Track B. Two routes could extend the framework there, and we flag both as *unvalidated by this paper*. The first is *model-based content adjustment*: replacing the exact-content stratum with a learned content representation and residualising against it, which converts a structural impossibility into a modelling assumption, the target then becomes identified only up to the adequacy of that content model, and a mis-specified model reintroduces exactly the exposure confound the exact-content design was built to remove. The second is *partial identification*: rather than a point estimate, report bounds on the target increment under explicit assumptions about how much of the between-user variation the unobserved content could explain, which is honest about sparse support but yields intervals that may be too wide to be useful. We have implemented and validated neither, and we do not know which is preferable; we note them so that "not estimable" is read as a statement about this design rather than about the question.

G Published-style diagnostic stress tests

To test whether the target-identification issue is specific to the abstract identity baseline, we implemented diagnostics motivated by recent personalized-representation analyses: cluster silhouette, variance along the leading user SVD component, pairwise top-k overlap, owner-retrieval AUC, displacement relative to norm-matched random directions, and within- versus between-user cosine gaps. The complete literature-to-diagnostic mapping and implementation are included with the released result package. Each diagnostic is evaluated both as normally reported and, where meaningful, against a matched null constructed from the same known-truth world.

Three kinds of threshold appear in this appendix and must not be conflated. (i) *Source-defined* criteria, stated by the paper that introduced the diagnostic, which we apply unchanged; (ii) *exploratory*



Figure 6: Common representation diagnostics are informative but not target-specific. Columns correspond to known-truth worlds; red indicates that the diagnostic fires. The column outlines encode *ground truth*: **solid-outlined** columns are worlds in which every user has the *same* mean target, so a firing there is a false positive, while **dotted-outlined** columns carry genuine mean heterogeneity, so a firing there is power and is desirable. The two panels encode the *calibration method*: the left panel scores each diagnostic in its usual reported form with no null, and the right panel scores the same diagnostic against a matched content-preserving null. Several diagnostics remain positive in equal-mean worlds, particularly under user-specific scalar dispersion, whereas the cross-fitted mean-target test remains calibrated. A positive diagnostic therefore need not identify mean heterogeneity even when the diagnostic is useful for its intended broader purpose. **Caution when counting red cells in the left panel:** only cluster silhouette, top-k Jaccard and displacement carry an absolute criterion stated by a source; the other three rows fall back to "above the permutation mean", which fires about half the time under any null, so their left-panel cells are not evidence either way.

criteria, used where no source states one, for which we fall back to "above the permutation mean", a bar that fires about half the time under *any* null and therefore carries no evidential weight; and (iii) *null-calibrated* tests, where the statistic is scored against a matched content-preserving null and a rejection is a genuine decision at $\alpha = .05$. Only (i) and (iii) support conclusions. **An asymmetry in the "as reported" column that must be read before the figure.** Only three of the six diagnostics come with an *absolute* firing criterion stated by a source: cluster silhouette (positive silhouette indicates clusters), top-k Jaccard ("values clustered near zero", i.e. below .5), and displacement relative to norm-matched random directions (outside the source's own two-sided control band). For the remaining three, leading-SVD user variance, owner-retrieval AUC, and the within- versus between-user cosine gap, no source states a threshold, so "as reported" necessarily falls back to *greater than the permutation mean*, which fires about half the time under *any* null by construction. **Those three "as reported" rates therefore carry no evidential weight, in this appendix or in Figure 6**, and we do not count them as diagnostics that fire. We retain them because their *calibrated* column is informative and because owner-retrieval AUC is our own calibrated owner-binding reference rather than a competitor's statistic, included to show what a calibrated diagnostic looks like.

A representative equal-mean scalar-variance world is especially revealing: cluster silhouette, leading-SVD variance, top-k Jaccard, and pairwise cosine-gap statistics all fire in every replicate as normally reported; after matched-null calibration, silhouette, SVD variance, and cosine gap still fire at rate 1.00, while the mean-target test fires at 0.035. In an exposure-only equal-mean world, matched-null calibration repairs most generic diagnostics, while a raw same-sample separation and excess-variance diagnostic still fire at 1.00; the cross-fitted mean target fires at 0.015. These results motivate target-specific rather than diagnostic-specific interpretation.

G.1 The same diagnostics on real activations under an equal-mean null

The known-truth experiments use synthetic data. To check whether the same diagnostic patterns appear with realistic activation geometry, we repeated the same diagnostic code on *real* cached MultiPref/Llama activations (layer 22, $K = 207$, $d = 4096$, pool $M = 21{,}061$, $n = 8$ per half),

using the empirical-null construction already employed for dispersion: permute the observation pool within stratum, which makes users exchangeable and so forces every user-level content-controlled mean to be equal, while preserving the real activation geometry, the real content exposure and the exact per-user counts. Every firing is therefore a false positive by construction. We ran 30 replicates $\times$ 100 permutations; this analysis is post-hoc and was added after internal review.

Among the three diagnostics that carry an absolute source-stated criterion, all three fire in **30 of 30** replicates as reported (95% CI $[.886, 1.000]$ each), and matched-null calibration reduces them to $1/30$, $0/30$ and $0/30$ for silhouette, Jaccard and displacement respectively. The synthetic finding therefore reproduces on real activations and is not an artifact of the generator. The remaining three diagnostics fire as reported at $11/30$, $16/30$ and $16/30$, which is the coin flip their fallback criterion implies and which we do not interpret.

One calibrated result does not follow the pattern. **Leading-SVD user variance still fires in** 5 **of** 30 **replicates after calibration** (16.7%, 95% CI $[.073, .336]$), an interval that excludes the nominal $.05$; the other five diagnostics are at or below nominal after calibration ($1/30$, $0/30$, $2/30$, $0/30$, $0/30$). We report this without a mechanism: we did not determine why the matched null is mismatched to that particular functional, and we do not infer a cause. The content-preserving reference corrects five of the six diagnostics examined on real activations, not all six. It is a useful control, not a general guarantee of calibration.

Three limits bound this experiment. It covers one dataset, one backbone, one layer and one depth, so it shows the synthetic result is not a generator artifact rather than surveying where the rates land elsewhere. The null is *stronger* than equal means: within-stratum permutation enforces full exchangeability and so equalises dispersion and every higher moment, meaning a diagnostic that fires only under dispersion heterogeneity with equal means would not be caught here. And 30 replicates resolve coarsely, the intervals distinguish "essentially always" from "essentially never" and little in between.

H TRACK-B MATCHED-K SUPPORT CONTROL

Tightening Community Alignment Track B from the released coarse grouping to exact comparison strata changes the broad identity decision, but also reduces the eligible cohort. To isolate the cohort-size channel, we repeatedly subsample the *coarse* Track-B population to exactly the $K = 127, 115, 86$ users retained by the exact analysis at $n = 2, 4, 8$. Across 500 subsamples per depth, coarse identity still rejects in 98.8%, 95.6%, and 89.6% of draws, whereas the exact-stratum identity test does not reject in the corresponding full exact cohorts ($p = .398, .264, .557$). The coarse identity effect size also remains roughly an order of magnitude larger than the exact one. Loss of K alone therefore does not reproduce the change.

This control does not isolate all consequences of exact matching. Tightening the stratum also removes 93.6% of observations as content singletons, reduces the permutable fraction from 1.000 to about 0.20, and collapses mean permutation-stratum size from roughly $1{,}010$ to 1.15. A secondary matched-K Track-A control confirms that exact-stratum identity testing is weak at these cohort sizes, rejecting only 7.8–20.0% of the time even where the full-K result is positive. Track B therefore illustrates how observation support changes the available evidence. Failure to reject after exact matching is not evidence that users are identical. The mean-target decision is identical under the coarse and exact Track-B definitions (fail at $n = 2$, pass at $n = 4, 8$).

I CROSS-SPLIT DEPENDENCE AND BLOCKED HISTORIES

Assumption 1 is directly stress-tested by adding correlation between the estimation errors in the two histories while holding the true user means equal. At $K = 100$ and $n = 8$ observations per half, the mean-target false-positive rate is 0.034 at $\rho = 0$, 0.980 at $\rho = .05$, and 1.00 for $\rho \geq .10$; 95% CI coverage falls from 0.944 to 0.020 already at $\rho = .05$. The induced bias, expressed in R units, is approximately ρ itself. Thus cross-estimation removes same-sample variance inflation only when the two errors are actually independent enough for the target being tested.

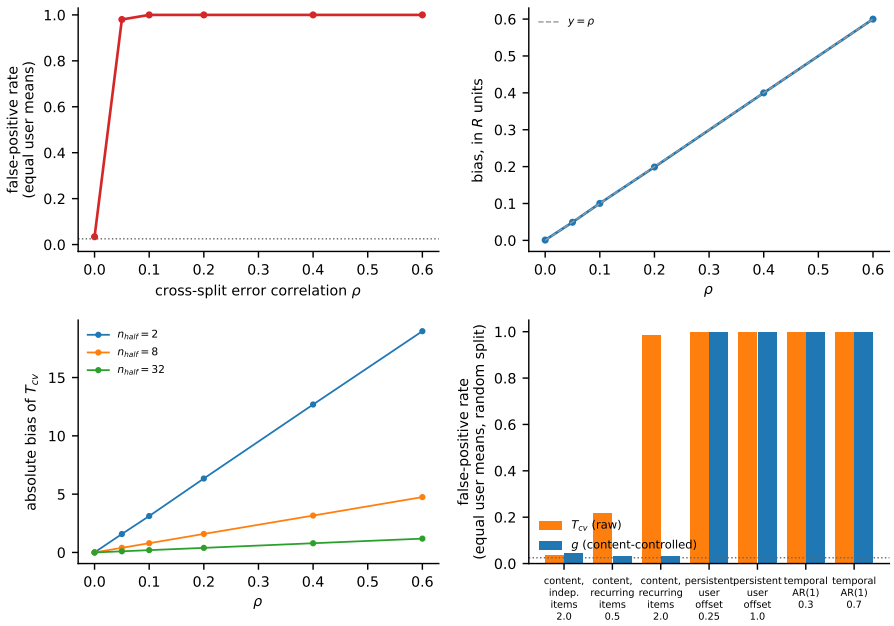

Figure 7: The cross-split independence condition has a sharp empirical consequence. Top-left: equal-target false-positive rate as cross-half error correlation increases. Top-right/bottom-left: induced bias is approximately linear in ρ and shrinks in raw units with deeper halves. Bottom-right: mechanistic null simulations under **random splitting only**, comparing the raw T_{cv} statistic with the content-controlled increment g; the dotted line marks the nominal level. The panel shows random splitting only; the blocked-split results appear numerically below.

What blocking does and does not repair. The same mechanistic simulation was run under both random and chronologically blocked splitting ($K = 100$, equal user means, so every rejection is a false positive), and the two arms separate sharply by *mechanism* rather than uniformly. Under short-range local dependence, blocking helps at moderate strength and stops helping as the dependence lengthens: with an AR(1) error at $\rho = 0.3$ the false-positive rate falls from 1.000 under random splitting to 0.387 under blocking, but at $\rho = 0.7$ it is 1.000 under both. Under a *persistent user-specific offset*, dependence that spans the whole history rather than adjacent observations, blocking does nothing at all: the rate is 1.000 under random and 1.000 under blocked splitting, at both tested magnitudes. Stable shared-content effects behave differently again: the raw statistic inflates under clustered content (0.217 random, 0.260 blocked at the weaker setting; 0.987 and 0.983 at the stronger), while the content-controlled increment g stays near nominal across both shared-content mechanisms (0.027–0.047 unclustered, 0.017–0.040 clustered), which is the separation the content-preserving reference is designed to produce. The same control does *not* rescue the temporal mechanisms, and is not meant to: under AR(1) errors g runs at 0.387–1.000 and under a persistent user offset at 1.000, because those violate cross-half independence rather than content exchangeability. Content control and dependence control are separate requirements, and only the first is a property of our reference.

Chronological blocking is a defence against short-range temporal dependence only, and the failure mode it does *not* touch, a user-specific offset persisting across the whole observation period, is precisely the one that would most directly inflate a mean-target interpretation. This is why our blocked MultiPref results are stated as conditional on Assumption 1 rather than as having established it, and why the buffered-gap arm, which widens the temporal separation further, is reported as a sensitivity analysis that fails our primary rule rather than as a repair.

Where real temporal information exists, we therefore test a blocked construction. MultiPref provides per-annotation timestamps. Under chronological halves, mean-target identification remains positive at $n = 8, 16, 32$ observations per half with $K = 207, 181, 133$ and permutation $p = .011, .006, .002$, respectively. The corresponding scale-free effects are $R = 2.62\%, 2.49\%, 4.56\%$. SFF does not provide equally clean timestamp structure for the same test; the available batch-blocked analysis is mixed and is treated as sensitivity rather than confirmation. Community Alignment carries no annotation timestamp, and its collection-wave variable does not vary within annotator; its Track-A

result therefore remains conditional on the random-split independence assumption rather than serving as blocked confirmation.

I.1 Operational support protocol

Support is not a single dataset-level percentage; it is a sequence of estimator-specific feasibility conditions, each of which is already enforced in code. We state them here as a protocol so that "the target is not estimable under this design" is a determinate verdict rather than a judgment call. For a target estimated at depth n (judgments *per split half*) on a user pool:

1. **Eligible repeated-content strata.** A content stratum is eligible iff at least two *distinct users* judged it ($k_s \geq 2$). Observations in singleton strata carry no within-content contrast, are dropped, and the dropped fraction is reported.
2. **Minimum observations per history estimate.** A user enters the mean target iff they hold at least $2n$ eligible observations, n for each disjoint half. The dispersion target additionally requires $n \geq 2$ per half, because its estimator is a pairwise U-statistic and is undefined at $n = 1$.
3. **Usable user population.** K is the number of users satisfying condition 2, plus, for held-out utility, at least 8 test judgments. K is thus a function of n: on MultiPref the same pool yields $K = 224, 219, 207, 181, 133, 63$ at $n = 2, 4, 8, 16, 32, 64$. The headline cohort is the $K = 133$ set fixed at the deepest confirmatory depth and held constant across shallower depths, so that depth comparisons are paired on identical users.
4. **Feasible label permutations.** The null requires that a permuted observation have somewhere to go: the reference is the fraction of the permutation pool lying in a stratum of size ≥ 2. When permutation is restricted to the sampled $2n$ observations this fraction can collapse, on SFF at $n = 2$ it is 0.122 at our split seed, and 0.098–0.122 across four seeds we checked , the permutation degenerates to the identity for the remaining observations, and the increment is then identically zero whatever the truth. We therefore permute over the full eligible pool and report the pool-level permutable fraction alongside every cell.
5. **Required cross-user content overlap.** For the dispersion target the leave-one-user-out reference must itself be estimable, so a user's observations must fall in strata whose co-member count $c_s = k_s - 1$ is large enough for the $O(1/c_s)$ exposure correction to be a correction rather than the estimate. We report the mean stratum size per cell, counted over the strata the estimator actually uses (those with $k_s \geq 2$, since singletons are dropped): 2.9 on MultiPref, 6.9 on SFF and 14.8 on Community Alignment Track A. Counted over *all* strata including singletons the same cells give 2.6, 6.9 and 14.6; we quote the former because it is the pool the residualisation sees. We treat c_s of order one as a warning, not a pass.
6. **Declared not estimable.** A target is *not estimable under the stated design* when condition 1, 2 or 4 fails, no eligible strata, too few users clearing $2n$, or a degenerate permutation reference. We report such cells as unsupported with the failing condition named, rather than as null results.

Structural feasibility is separate from statistical power. Conditions 1–5 decide whether an estimate *exists*; they say nothing about whether a true effect would be detected. A cell can be fully supported and still be uninformative because K is small or reliability is low, and we report K and R for every cell so the two can be read apart. Where a cell clears the protocol but has low power we say the target was *not detected*, never that it is homogeneous.

I.2 Reliability indexing and model-based projections

In $R(n)$, n is the number of judgments *per split half*; a user contributing $n_{\text{total}} = 2n$ judgments supplies two half-estimates, so any figure or table indexed by n_{total} reports $R(n_{\text{total}}/2)$. Every R we quote is measured at a depth we actually ran. A value obtained for an unrun depth by applying the Spearman–Brown aggregation law, in particular a full-history projection, is *model-based under the stylized random-effects decomposition, not an empirically measured reliability*, and we label it as such wherever one is given.

I.3 SCOPE OF THE WITHIN-STRATUM PERMUTATION TEST

The within-stratum identity permutation makes user labels exchangeable given content. It therefore tests $P(z \mid U = u, S = s) = P(z \mid U = v, S = s)$, which is strictly stronger than the mean-target null $\mathbb{E}[z \mid U = u, S = s] = \mathbb{E}[z \mid U = v, S = s]$ that our estimand concerns: a population with equal means but unequal within-user dispersion satisfies the second and violates the first. The gap is not hypothetical, because our dispersion target reports exactly that kind of heterogeneity.

We measured it. In controlled worlds with *identical* true user means ($K = 120$, $n = 8$ per half, 60 strata, 1,000 replicates per world, $\alpha = .05$), the permutation p-value fires at .047 when within-user dispersion is homogeneous but at .084–.120 once users differ in dispersion, with assortative content exposure, unequal observation counts and heavy tails each leaving it near .09. Over the same replicates the one-sided user bootstrap on the per-user increment holds at .041–.057. Our preregistered Stage-2 rule declares the target detected when $Q_{.025}(\{\bar{g}^*\}) - 1.645\,\widehat{\mathrm{se}}_{\mathrm{MC}} > 0$, where $Q_{.025}(\{\bar{g}^*\})$ is the 2.5th percentile of the user-bootstrap distribution of the mean per-user increment and $\widehat{\mathrm{se}}_{\mathrm{MC}}$ accounts for Monte Carlo uncertainty in the estimated content-preserving reference. The rule runs conservatively at .011–.033, at a power cost (e.g. .51 versus .78 at a small true effect). Two-sided interval coverage is .933–.953 against a nominal .95. The subsample-scope and full-pool permutations behave alike.

Consequently we report the permutation p-value only as a test of exchangeability and rest every mean-target decision on the preregistered bootstrap rule, which is what the frozen analysis plan already specified. This matters most where the two disagree, which among our analyses is only the buffered-gap arm (Section I.4).

I.4 BUFFERED-GAP TEMPORAL SENSITIVITY

This post-hoc analysis, added after internal review, assesses sensitivity to short-range temporal dependence in MultiPref. We retain the fixed $K = 133$ cohort and the same three representation spaces. Each user's timestamped build observations are divided into balanced chronological thirds without splitting identical-timestamp groups; a fixed sampling rule selects $n = 16$ observations per third. We compare early–middle, middle–late, and buffered early–late histories (middle third omitted), alongside the original chronological $16 + 16$ split. All arms use the same users, model layers, 16 judgments per half, full-pool within-content reference, and user-level bootstrap. The selected observations and content composition necessarily vary between arms. Each model uses 1,000 permutations and 10,000 bootstrap resamples. The original chronological target increment, $R(16)$, and permutation p reproduce the saved Experiment-I results for all three models.

The median across users of the *mean between-history separation* is 5.36, 4.56, 11.06, and 8.00 days for early–middle, middle–late, buffered early–late, and original chronological splits, respectively. Their median *boundary gaps* are 0.02, 0.02, 5.49, and 0.07 days. Selected shared-content fractions are .430, .435, .402, and .391, respectively; the separate full-pool reference retains approximately .946 shared-content support. The buffered comparisons therefore increase separation but do not hold the selected content fixed.

Table 4 gives all arm-level results. Buffered increments remain positive in point estimate but smaller than the original chronological arm, and **the buffered arm fails our prespecified decision rule of Section 3 in all three backbones**: the Monte-Carlo-adjusted one-sided lower bounds are $-.0047$, $-.0477$ and $-.0019$ for Llama, Qwen and Mistral, against $+.0280$, $+.8485$ and $+.0133$ for the original chronological arm, which passes in all three. The two-sided 95% user-bootstrap intervals cross zero in the same three cells, consistently. The *unadjusted, one-sided* permutation p-values for the buffered arm are $(.037, .016, .024)$, but by Section I.3 a permutation p-value tests exchangeability rather than equality of mean targets and is anti-conservative precisely when users differ in dispersion; we therefore do not treat these three values as evidence that the buffered effect is nonzero, and we report them only for comparability with the other arms. We also apply no multiplicity adjustment across the twelve arm-by-model cells: the preregistered families in this paper are within-model, and fitting a new family to a post-hoc sensitivity analysis after seeing its outcome would itself be a researcher degree of freedom. Under Holm across the twelve permutation p-values the three buffered cells would not survive, which we note rather than adopt. No arm demonstrates zero cross-history error covariance.

Table 4: Post-hoc MultiPref buffered-gap sensitivity ($K = 133$, $n = 16$ per half). Effects $\hat{\Delta}$ use each model's own representation units and are not comparable across models. Intervals are two-sided 95% user-bootstrap intervals. "Prespec. bound" is the Monte-Carlo-adjusted one-sided lower bound that constitutes our prespecified decision rule (Section 3); the target is identified only when it exceeds zero. Permutation p-values are one-sided, unadjusted, and test exchangeability rather than equality of mean targets (Section I.3). The early–late arm omits the middle third.

Model	Split	$\hat{\Delta}$ [95% CI]	Prespec. bound	Perm. p	$R(16)$
Llama	Early–middle	$+.0379$ $[-.0059, +.0859]$	$-.0069$ (fail)	.022	2.02%
	Middle–late	$+.0846$ $[+.0400, +.1341]$	$+.0389$ (pass)	.001	4.49%
	Early–late buffered	$+.0380$ $[-.0037, +.0822]$	$\mathbf{-.0047}$ **(fail)**	.037	1.95%
	Original chronological	$+.0742$ $[+.0291, +.1222]$	$+.0280$ (pass)	.001	3.77%
Qwen	Early–middle	$+1.5319$ $[+.3738, +2.8769]$	$+.3483$ (pass)	.002	3.28%
	Middle–late	$+2.2714$ $[+1.1522, +3.4582]$	$+1.1261$ (pass)	.001	4.82%
	Early–late buffered	$+1.2454$ $[-.0163, +2.6323]$	$\mathbf{-.0477}$ **(fail)**	.016	2.58%
	Original chronological	$+2.0696$ $[+.8765, +3.3782]$	$+.8485$ (pass)	.001	4.22%
Mistral	Early–middle	$+.0130$ $[-.0002, +.0278]$	$-.0005$ (fail)	.011	2.00%
	Middle–late	$+.0285$ $[+.0140, +.0445]$	$+.0137$ (pass)	.001	4.34%
	Early–late buffered	$+.0122$ $[-.0015, +.0274]$	$\mathbf{-.0019}$ **(fail)**	.024	1.82%
	Original chronological	$+.0271$ $[+.0135, +.0417]$	$+.0133$ (pass)	.001	4.00%

Table 5: User-paired buffered-minus-comparator increments. Differences and two-sided 95% bootstrap intervals use each model's representation units. All comparisons are post-hoc and unadjusted.

Model	Comparator	Difference	95% CI
Llama	Early–middle	$+.0001$	$[-.0538, +.0544]$
	Middle–late	$-.0467$	$[-.1041, +.0092]$
	Original chronological	$-.0362$	$[-.0913, +.0169]$
Qwen	Early–middle	$-.2865$	$[-1.6987, +1.1717]$
	Middle–late	-1.0261	$[-2.5701, +.4478]$
	Original chronological	$-.8242$	$[-2.4170, +.7562]$
Mistral	Early–middle	$-.0008$	$[-.0169, +.0153]$
	Middle–late	$-.0162$	$[-.0331, -.0001]$
	Original chronological	$-.0149$	$[-.0320, +.0017]$

The user-paired contrasts in Table 5 separate attenuation in point estimates from evidence for a difference. All three buffered-minus-original intervals include zero; only the Mistral buffered-minus-middle–late interval narrowly excludes zero before any multiplicity adjustment. The larger gap can change transient dependence, the target over time, and content composition simultaneously, while long-lived user-specific nuisance may survive the buffer. Consequently, this is sensitivity evidence, not a causal intervention, an estimate of dependence inflation, or proof of preference drift.

J KNOWN-TRUTH POPULATION-SIZE × HISTORY PHASE DIAGRAM

The stylized argument in Proposition 2 predicts a low-signal regime in which population heterogeneity can be detected using many users while per-user personalization remains difficult. We test this in a controlled generator over $K \in \{25, 50, 100, 200, 500, 1000\}$ and history depths $n_{\text{total}} \in \{4, 8, 16, 32, 64, 128\}$. At the low-signal setting, a broad diagonal band appears in which identification power is high while the personalized-utility test has low power; strengthening the latent signal by $3.75\times$ removes most of this band. At signal levels matched to the observed SFF/MultiPref scale, utility power is only about 0.38–0.39 in the corresponding controlled cells, underscoring that a non-significant utility contrast is a power-bounded statement rather than proof of zero value.

For comparison, the earlier empirical K-sweep fixes Community Alignment, model, layer, and $n = 8$ while varying only the target-user population. Detection rises from 0.18 at $K = 25$ to 0.59 at 50, 0.91 at 100, and 1.00 at 200 while a fixed-reference-pool arm keeps the estimated target increment approximately stable. This is consistent with the population-power side of the stylized result but is not itself a sample-complexity theorem.

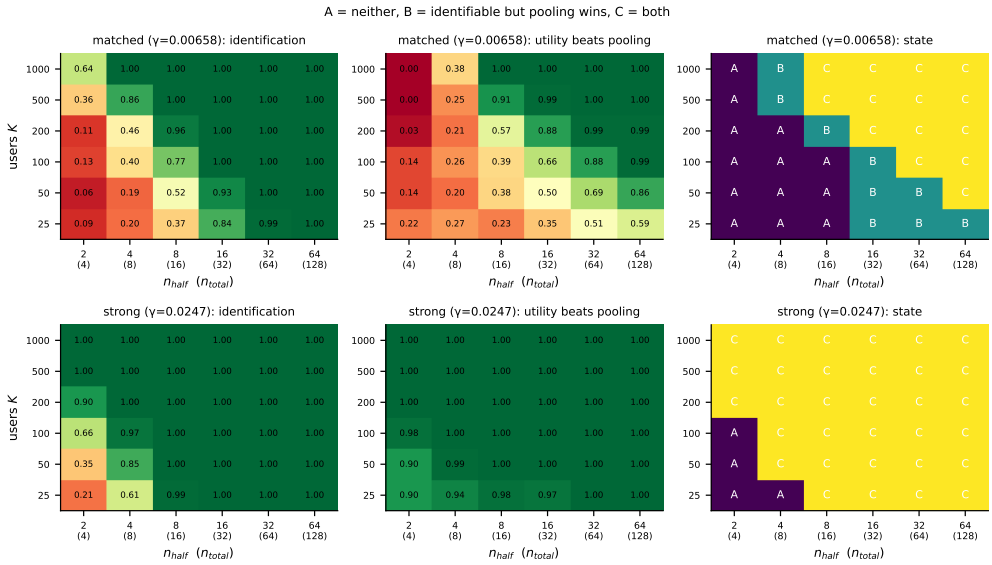

Figure 8: Known-truth $K \times n$ phase diagram. The top row uses a low-signal generator and the bottom row a $3.75\times$ stronger signal. Left/middle panels show target-identification and utility-test power; right panels classify each cell as neither (A), identifiable without detected utility (B), or both (C). The B region is wide only in the low-signal regime.

K Same-design depth sweeps and three-model MultiPref replication

The SFF confirmatory sweep fixes $K = 54$ users with at least 128 build and 8 test judgments, then varies only total build history $n_{\text{total}} \in \{4, 8, 16, 32, 64, 128\}$. Identification consumes exactly the same history by splitting it into two halves. The primary utility estimator uses user-level cross-fitted hyperparameter selection and is compared with leave-target-out pooling. At $n_{\text{total}} = 128$, all three models pass target identification after their preregistered within-model depth correction while no model establishes utility over pooling. Qwen uniquely establishes utility at $n_{\text{total}} = 16$ ($+0.01229[+0.00414, +0.02132]$; raw $p = .0012$, Holm-adjusted $p = .0072$ across the preregistered six-depth Qwen utility family) while the mean-target test is null ($R = -0.00107$, permutation $p = .527$). The adjacent $n_{\text{total}} = 32$ interval excludes zero before multiplicity correction but does not survive the same family ($p_{\text{Holm}} = .065$).

MultiPref uses a fixed $K = 133$ cohort at $n_{\text{total}} \in \{4, 8, 16, 32, 64\}$. The two MultiPref cohort sizes that appear in this paper come from different, explicitly different, inclusion rules: the identification cohort $K = 133$ is the set of users with at least $2n = 64$ eligible build observations (the deepest confirmatory depth), held fixed across shallower depths; the full-history utility cohort $K = 183$ is the set of users present in both build and test splits with at least 8 build observations and at least 8 held-out test judgments. Neither is a subset selection made after seeing results. The original Llama replication was followed by a preregistered three-model extension frozen before the Qwen/Mistral activations existed. The headline target analysis uses chronological blocks based on per-annotation timestamps. At $n_{\text{total}} = 16, 32, 64$, all nine model-by-depth cells are positive under the preregistered within-model Holm families. The corresponding R values are $(2.74, 3.77, 4.56)\%$ for Llama, $(2.82, 4.22, 5.14)\%$ for Qwen, and $(2.98, 4.00, 4.47)\%$ for Mistral. None of the 15 model-by-depth utility cells establishes a gain over leave-target-out pooling. A post-hoc Holm sensitivity across all nine headline identification cells also leaves all nine significant; it is reported only as a robustness check and does not replace the preregistered families.

L Layer sensitivity

The primary Llama layer (22; relative depth 0.688) was frozen before the headline analyses. As post-hoc robustness, we repeated the MultiPref $n_{\text{total}} = 64$ chronological target analysis at layers 16 and 27 using activations that had already been extracted; all three tested layers are reported. The target is

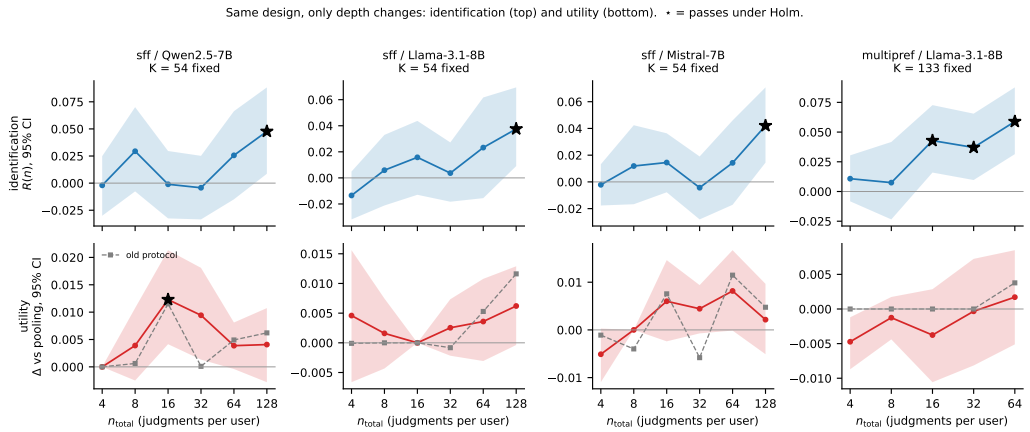

Figure 9: Same-design depth sweeps: three SFF backbones and MultiPref/Llama. Identification (top row) and utility (bottom row) use the same fixed users and the same total history budget at each depth, so the two endpoints are never given different amounts of history. **Columns 1–3** are SFF at $K = 54$ fixed (Qwen, Llama, Mistral); **column 4 is MultiPref/Llama at $K = 133$ fixed** and is included here for the same-design comparison, not as part of the SFF sweep. A star marks a cell passing under its preregistered within-model Holm family. On SFF, Qwen shows utility without mean-target identification at $n_{\text{total}} = 16$, whereas all three SFF backbones show mean-target identification without an established utility gain at $n_{\text{total}} = 128$. The MultiPref column shows identification at every depth from $n_{\text{total}} = 16$ with no utility cell clearing pooling. The grey dashed series in the bottom row is the older half-split utility protocol, retained for comparison.

Table 6: Community Alignment Track-A mean-target replication. $R(8)$ is the scale-free controlled reproducible share.

Model	Layer	K	Mean-target increment [95% CI]	$R(8)$
Llama-3.1-8B	22	1,929	$+0.2035$ $[+0.1775, +0.2313]$	8.8%
Qwen2.5-7B	20	1,929	$+6.3622$ $[+5.6103, +7.1579]$	11.5%
Mistral-7B-v0.3	22	1,929	$+0.0606$ $[+0.0532, +0.0683]$	8.5%

identified at every layer: increments are $+0.0167[+0.0071, +0.0270]$, $+0.0460[+0.0187, +0.0749]$, and $+0.1244[+0.0471, +0.2073]$ for layers 16, 22, and 27. Although the raw hidden-space increment changes by roughly sevenfold, R is stable at 0.050, 0.046, and 0.054. This supports the intended use of R as a scale-free finite-history target-reliability summary and shows that the MultiPref conclusion is not unique to layer 22.

M Large-population Community Alignment replication

On Community Alignment Track A, the same $K = 1929$ annotators are evaluated at $n = 8$ observations per half under complete shared-content support on the identification split. The target-identification pattern appears across all three backbones, although raw increments are not comparable across representation scales. Because CA has no annotation timestamp and no within-annotator collection-wave variation, these random-half results are replication under Assumption 1, not dependence-robust blocked confirmation.

N Utility baselines and full-history results

The pooled baseline is leave-target-out: for evaluated user u, the population vector is constructed from other users only. Scalar partial pooling uses

$$\hat{v}_u(\lambda) = \lambda \hat{v}_u + (1 - \lambda)\bar{v}_{-u}, \tag{20}$$

with λ chosen on calibration users disjoint from the evaluation users. The hierarchical estimator learns a low-rank population subspace and prior from separate users, selects rank and regularization on calibration users' held-out judgments, and applies the resulting shrinkage rule to the target user without using that user's held-out labels. Across the prespecified fixed-depth grid, the hierarchical implementation beats pooling in $0/15$ human cells and $9/9$ synthetic positive controls.

Table 7: Full-history utility grid. Effects are held-out accuracy minus leave-target-out pooling. Shrinkage λ is selected on a disjoint user split. Boldface marks an interval excluding zero; this table is descriptive and is separate from the preregistered six-depth SFF family reported in Section K. The Track-B cell is positive but not target-specific because held-out shared-content support is only 0.025.

Dataset	Model	K	Median hist.	Local [95% CI]	Shrinkage $\lambda^\star$ [95% CI]	Hierarchical [95% CI]
CA-A	Llama	567	20	$-.0095[-.0196, +.0007]$	$.4:\ -.0044[-.0128, +.0040]$	$-.0014[-.0055, +.0024]$
CA-A	Mistral	567	20	$-.0095[-.0201, +.0012]$	$.4:\ -.0011[-.0097, +.0073]$	$+.0019[-.0040, +.0073]$
CA-B†	Llama	400	104	$+.0156[+.0023, +.0290]$	$.5:\ +.0221[+.0092, +.0354]$	$+.0289[+.0175, +.0399]$
MultiPref	Llama	183	100	$-.0121[-.0270, +.0026]$	$.7:\ -.0099[-.0285, +.0084]$	$-.0020[-.0090, +.0050]$
SFF	Llama	68	296	$-.0101[-.0348, +.0108]$	$0:\ 0[0, 0]$	$-.0066[-.0141, +.0007]$
SFF	Qwen	68	296	$+.0138[-.0107, +.0356]$	$.2:\ \mathbf{+.0224[+.0081, +.0395]}$	$+.0070[-.0066, +.0220]$
SFF	Mistral	68	296	$+.0171[-.0033, +.0347]$	$1:\ +.0192[-.0010, +.0426]$	$+.0026[-.0054, +.0107]$

†Diagnostic only: the held-out shared-content rate is 0.025.

O PRIVILEGED-HISTORY DIAGNOSTIC

The first privileged-history analysis was specified before its results were inspected but used only a matched subset of held-out history and was subsequently recognized as shallower than the full-history estimator. After inspecting that first pass, a second arm was added. The added arm estimates each user's direction using all build-split observations plus one disjoint half of the held-out block, then scores on the other half. It therefore has strictly more user information than the deployable estimators but is not a mathematical upper bound over possible algorithms. Because it also changes the scoring split, its values are not numerically comparable with the primary all-held-out full-history effects in Table 7. The contrast that the privileged arm is actually about, privileged minus *matched local*, paired on the same users and scored on the same held-out half, is computable from the per-user accuracies, and we report it for *all seven* cells rather than leaving the two arms to be compared by eye: $+.0038\,[-.0071, +.0153]$ on MultiPref/Llama ($K = 183$); $+.0006\,[-.0084, +.0098]$, $+.0112\,[+.0008, +.0233]$ and $-.0037\,[-.0163, +.0088]$ on SFF for Llama, Qwen and Mistral ($K = 68$); $+.0071\,[-.0015, +.0157]$ and $+.0192\,[+.0099, +.0284]$ on CA-A for Llama and Mistral ($K = 567$); and $+.0165\,[+.0083, +.0246]$ on Track B ($K = 400$). **Three of the seven intervals exclude zero**, CA-A/Mistral, Track B and SFF/Qwen, and under Holm correction across the seven cells two survive (CA-A/Mistral and Track B, both $p_{\text{Holm}} = .0007$; SFF/Qwen reaches $p_{\text{Holm}} = .075$).

Which content support is and is not present matters here, and must be stated per split. Community Alignment Track A has a shared-content rate of 1.000 on the *build* split used for target identification, and 0.000 on its *held-out* split: no test stratum is shared between users (maximum users per test stratum is 1, and no stratum appears in both splits). Track B's held-out shared-content rate is 0.025. So neither CA-A nor Track B supports a *content-controlled* held-out comparison across users, and the privileged-history contrast is not one. It does not need to be: privileged minus matched local is a *paired within-user* accuracy difference, computed for each user on that user's own held-out half with both arms scored on the identical items, so it is valid whether or not other users saw the same content. What the absent held-out content support does forbid is reading these gains as target-specific, they say a richer estimator scores better on that user's own items, not that a named user-level functional has been identified out of sample.

The corrected reading is therefore weaker than either extreme: privileged history buys a paired held-out gain that survives multiplicity in two representation spaces (CA-A/Mistral and Track B), shows an unadjusted interval excluding zero in a third that does not survive (SFF/Qwen, $p_{\text{Holm}} = .075$), and shows nothing in the remaining four on the same data, and because the privileged arm is not an upper bound over algorithms, and because it compares estimators' accuracy rather than testing a functional against a content-preserving reference, we cannot say what the gap implies about achievable personalization or about any target, in the cells with fully shared held-out content (SFF, MultiPref) no less than in those without.

Distinguishing unadjusted intervals from multiplicity-adjusted decisions matters here. Three of the seven paired privileged-minus-matched intervals exclude zero without adjustment (CA-A/Mistral, Track B, SFF/Qwen), but under Holm across the seven cells only **two** are decisions: CA-A/Mistral and Track B ($p_{\text{Holm}} = .0007$ each). **SFF/Qwen does not survive** ($p_{\text{Holm}} = .075$) and we do not call it significant. Read at the level of decisions, the endpoint responds to additional user information in two of seven cells, even where the interval against pooling still crosses zero. On synthetic personalization-positive worlds, the same construction clears pooling in $7/9$ cells, while using the latent true user

Table 8: Post-hoc privileged-history diagnostic. Accuracy relative to pooling for matched local and privileged-history estimators, scored on the same held-out half. The paired privileged-minus-local comparisons and content-support limitations are discussed in the text.

Dataset	Model	K	Build hist. median	Matched local vs pool [95% CI]	Privileged vs pool [95% CI]
CA-A	Llama	567	20	$-.0148[-.0271, -.0022]$	$-.0077[-.0201, +.0046]$
CA-A	Mistral	567	20	$-.0108[-.0249, +.0031]$	$+.0084[-.0046, +.0214]$
CA-B†	Llama	400	104	$+.0129[-.0035, +.0294]$	$+.0294[+.0129, +.0456]$
MultiPref	Llama	183	100	$-.0031[-.0239, +.0172]$	$+.0007[-.0197, +.0208]$
SFF	Llama	68	296	$-.0189[-.0495, +.0113]$	$-.0183[-.0480, +.0111]$
SFF	Qwen	68	296	$+.0098[-.0156, +.0335]$	$+.0209[-.0039, +.0427]$
SFF	Mistral	68	296	$-.0012[-.0292, +.0224]$	$-.0049[-.0335, +.0187]$

†Held-out shared-content rate 0.025; not target-specific.

direction clears pooling in $9/9$. These controls motivate the diagnostic but do not convert it into an oracle bound.

P INTERVENTIONAL EVALUATION

Interventions are reported as separate behavioural evidence rather than as target-identification tests. Three corpora use different pooled-direction definitions and endpoint scales, so signs and qualitative null/positive patterns may be compared but effect magnitudes are not pooled or ranked across corpora.

MultiPref / Llama (primary clean intervention). At $n_{\text{total}} = 64$, 40 calibration users select steering strength from the frozen grid $\{0.25, 0.5, 1.0\}$ and 93 untouched users form the evaluation set; $\alpha^\star = 1.0$. The pooled arm is leave-target-out, matching the predictive-utility comparison. Pooled steering improves over unsteered by $+0.002761[+0.002078, +0.003431]$. Owner steering beats norm-matched user-span random by $+0.002704[+0.001684, +0.003922]$ and isotropic random by $+0.002720[+0.001594, +0.004009]$, with both bootstrap and signed-rank sensitivity agreeing. Owner minus pooled is $-0.000212[-0.001170, +0.000840]$. Owner minus stranger is positive under the prespecified mean-based bootstrap, $+0.001257[+0.000233, +0.002421]$, but not under the signed-rank sensitivity ($p = .112$); the per-user distribution is strongly right-skewed, so we treat owner-specific superiority as sensitivity-dependent rather than established. The evaluation is partially shared-content: 1,846 scored pairs cover 1,298 distinct comparisons, and 54.1% of scored pairs occur on a comparison scored by another evaluation recipient.

Regularized (partially pooled) steering, post hoc. Because owner-minus-pooled is null above, we asked whether *partial* pooling recovers a gain that neither extreme shows. For recipient u we steer with $v_u(\lambda) = \lambda v_u + (1 - \lambda)\bar{v}_{-u}$, where $\bar{v}_{-u}$ is the same leave-target-out pooled direction, so $\lambda = 1$ reproduces the owner arm and $\lambda = 0$ the pooled arm exactly. We reuse the *existing* 40/93 calibration–evaluation split, select $(\lambda^\star, \alpha^\star)$ over $\lambda \in \{0, .25, .5, .75, 1\} \times \alpha \in \{.25, .5, 1\}$ on calibration recipients only by mean gain over pooled at matched α, and freeze before any evaluation recipient is scored. The selected value is $\lambda^\star = 0$: on calibration users *every* non-zero shrinkage has a negative mean gain over pooling (between $-.00022$ and $-.00035$ across the grid). On the 93 held-out recipients at $\alpha^\star = .25$, owner minus pooled is $-0.000102\,[-0.000348, +0.000148]$. Partial pooling therefore does not recover an owner-specific steering gain, and the shrinkage that calibration selects is the one that discards the user vector entirely. This analysis is post hoc, its two hyperparameters were selected on calibration users, and the intervals are not adjusted for that selection; we report it as a negative result rather than as a tuned estimate.

Shared-content steering sensitivity, post hoc. The intervention endpoint is only partially content-controlled, so we repeated the contrast on the exact-comparison subset. Using per-pair steered margins, a scored pair is retained when its `comparison_id` is also scored by at least one *other* evaluation recipient: 1,008 of 1,846 pairs (54.6%) across all 93 recipients, none of whom is dropped. On that subset owner minus pooled is $-0.0000996\,[-0.000446, +0.000239]$, against $-0.000102\,[-0.000348, +0.000148]$ on the full set. Restricting to shared content therefore does not reveal an effect that the partially-shared endpoint was masking; the null is the same in both, and the retained-subset interval is wider, as expected from roughly half the pairs. (The 54.6% here is

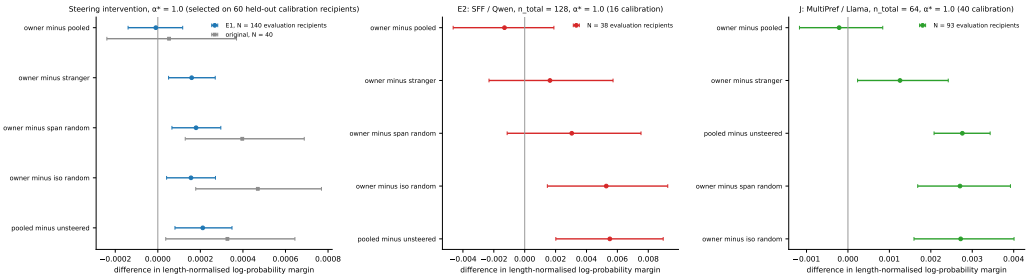

Figure 10: Steering intervention evidence across three corpora. Error bars are 95% user-bootstrap intervals. **The "pooled" baseline is not the same object in every panel and the panels must not be read as a common comparison.** The MultiPref panel uses a *leave-target-out* pool, the mean of the other $K - 1$ user directions, excluding the evaluated user, so "owner minus pooled" contrasts a user's own direction against a pool that never saw them. The E1 (Community Alignment Track B) and E2 (SFF) panels use a *self-inclusive grand-mean* pool, which contains the evaluated user's own direction and is therefore biased toward the owner arm by construction; an owner-minus-pooled value there is not comparable with MultiPref's, and a null there is a weaker statement. Endpoint scales also differ (corpus-specific log-probability margins), so only qualitative patterns should be compared across panels and no magnitude should be ranked. MultiPref is the cleanest intervention on both counts: leave-target-out pooling, and a target result that survives chronological blocking.

computed over the 93 evaluation recipients of this run; the 54.1% quoted above uses the support audit's slightly different denominator.)

Community Alignment and SFF. The scaled Community Alignment Track-B intervention uses 60 calibration and 140 evaluation recipients. Owner steering is positive relative to stranger and matched random controls, pooled steering beats unsteered, and owner minus pooled is tightly null; because Track B has poor exact content support, this is owner-binding/interventional evidence rather than a target-specific confirmation. The SFF/Qwen intervention uses 16 calibration and 38 evaluation recipients at $n_{\text{total}} = 128$; pooled steering beats unsteered and owner beats isotropic random, while owner minus pooled is null and other owner-specific contrasts are inconclusive. E1/E2 use a self-inclusive grand-mean pooled direction, unlike MultiPref's leave-target-out pool, and the corpus-specific log-probability margins differ in scale; we therefore do not compare magnitudes across these experiments.

Q PERFIT IMPLEMENTATION DETAILS

PerFit (Liu et al., 2026) is run at commit `a3fbd8bcd087eb85436e876042db60ab79e00101`. Its collective stage is trained once and the user-specific stage twice per evaluated user on disjoint history halves. The 24 GB launcher uses gradient accumulation while preserving effective batch size and optimizer updates. Content-dependent block outputs are cached for target extraction, and a unit test verifies the recomputed local contribution against PerFit's forward pass. A pre-test support audit leaves LaMP-2M as the usable repeated-content case. Stages D–E use deterministic held-out candidate-tag scoring rather than PerFit's published free-generation metric, so the values are not directly comparable.

Q.1 CANDIDATE-TAG SCORING (STAGES D–E)

Scoring is deterministic: no sampling and no free generation. The candidate set is the fixed $|\mathcal{T}| = 15$ LaMP-2M tag strings, all scored in one padded batch against a shared prompt, so the intervention positions are identical across candidates. For a tag τ tokenised as $(\tau_1, \ldots, \tau_{L_\tau})$ and prompt $\rho(x)$, the score is the *mean per-token* log-probability, i.e. normalised by tag token length, and the prediction is its arg-max:

$$\mathrm{sc}_\theta(\tau \mid x) = \frac{1}{L_\tau} \sum_{t=1}^{L_\tau} \log p_\theta\big(\tau_t \mid \rho(x), \tau_{<t}\big), \qquad \hat{\tau}_\theta(x) = \arg\max_{\tau \in \mathcal{T}} \mathrm{sc}_\theta(\tau \mid x). \tag{21}$$

With θ ranging over the frozen stage-1 collective, the user's two half-models, and a stranger's half-A model under the fixed deterministic pairing $\sigma(u_i) = u_{(i+7) \bmod K}$, per-user accuracy on the user's

own shared-content items is $\mathrm{acc}_\theta(u) = |\mathcal{I}_u|^{-1} \sum_{x \in \mathcal{I}_u} \mathbf{1}[\hat{\tau}_\theta(x) = \tau^\star(x)]$, and with $\mathrm{acc}_{\mathrm{own}}(u) = \frac{1}{2}[\mathrm{acc}_{uA}(u) + \mathrm{acc}_{uB}(u)]$ the two reported margins are the user-bootstrapped means of $\mathrm{acc}_{\mathrm{own}}(u) - \mathrm{acc}_{\sigma(u)A}(u)$ and $\mathrm{acc}_{\mathrm{own}}(u) - \mathrm{acc}_{\mathrm{coll}}(u)$ over the $K = 94$ users, giving $+.0444\,[+.0249, +.0667]$ and $+.0265\,[+.0131, +.0409]$ (mean accuracies .551, .506 and .524 for own, stranger and collective).

Q.2 The owner-matching reference, and its null-dependence

The owner-matching statistic is a row-wise rank score: with $\hat{A}_u, \hat{B}_u$ the grand-mean-centred, unit-normalised half-means and $S_{uv} = \langle \hat{A}_u, \hat{B}_v \rangle$, $\mathrm{AUC} = K^{-1} \sum_u (K-1)^{-1} \sum_{v \neq u} \mathbf{1}[S_{uu} > S_{uv}]$, for which **chance is** .500. The reported comparator .726 is *not* chance: it is the mean of this statistic over 1,000 within-item user-label permutations (SD .020, 95% range $[.686, .766]$), i.e. the owner-matching score attributable to shared item composition alone, because a within-item permutation maps the set of users holding an item onto itself and so preserves each user's exact item set. The observed value is .793, with a user-bootstrap 95% interval of $[.738, .825]$; the increment is $+.068\,[+.012, +.099]$, which excludes zero, and the permutation p-value is .003. We note two things that the level .793 alone would hide: the permutation null is wide enough that its maximum over 1,000 draws (.809) exceeds the observed value, so most of the level is item composition rather than user binding; and the result is *null-construction-dependent*. Under the alternative coupled null, in which the two halves are permuted together rather than independently, the reference rises to .880 and the test does **not** reject ($p = 1.0$). We report both and rest the external case on the mean target rather than on owner matching.

Q.3 Cross-half reproducibility, and why it is not a target reliability

We can compute the ratio $4.380/9.402 = 0.466$, and we report it, but **as a descriptive cross-half reproducibility ratio, not as a validated fraction of true-target variance**. This subsection sets out exactly which assumptions that stronger reading would need and which of them the PerFit design can and cannot supply.

The decomposition that matters. Write the two centred half-estimates as

$$A_u = \theta_u + b_u + \varepsilon_{uA}, \qquad B_u = \theta_u + b_u + \varepsilon_{uB}, \tag{22}$$

where θ_u is the intended population target, b_u is any *user-specific* component shared by both halves that is not part of that target, and $\varepsilon_{u\cdot}$ is half-specific error. The cross-fitted statistic has $\mathbb{E}[\langle A_u, B_u \rangle] = \|\theta_u + b_u\|^2$ whenever $\varepsilon_{uA} \perp \varepsilon_{uB}$: it is blind to the split between θ_u and b_u, because both are user-specific and both persist across halves.

What the cosine diagnostics do and do not settle. Before centring, within-user $\cos(A_u, B_u) = .982$ and cross-user $\cos(A_u, B_w) = .955$; after centring across users these become .463 and .022. It would be wrong to read high raw cosine as proof of dependence, almost all of it is a component common to every user (the frozen stage-1 collective, 92.5% of raw energy) which the centring removes. It would be equally wrong to read the centred cross-user value of .022 as establishing independence. That number says only that no *user-independent* component survives centring. b_u is by construction user-specific, so it contributes to the within-user value and not to the cross-user value, and **neither cosine can separate θ_u from b_u**. No diagnostic computable from the released artifacts can.

Is there a concrete candidate for b_u here? Yes. Two features of this design generate exactly such a term. Both half-models are initialised from the same frozen collective under the same seed and differ only in which profile rows they see, so any user-specific basin the optimisation falls into, driven by, say, the length or surface statistics of that user's profile rather than their preferences, is common to both halves. And both halves are read out over the *same* item set for each user, so any user-specific interaction between a trained model and its own items enters A_u and B_u identically. The content-preserving null does not remove these: a within-item permutation destroys user identity, so anything user-specific, target or not, is absent from the null and therefore survives into the increment.

The four assumptions, and their status. Reading 0.466 as a target reliability requires all of:

(A1) $\varepsilon_{uA} \perp \varepsilon_{uB}$ given the user. **Partly supported, not verified.** The halves train on disjoint profile rows, and centred cross-user cosine of .022 rules out a shared user-independent component. Shared initialisation, seed and collective remain, and they bear on user-specific dependence, which that diagnostic cannot see.

(A2) $b_u = 0$, i.e. every user-specific reproducible component is target. **Untestable with the released artifacts**, and §(above) gives two concrete mechanisms that would violate it.

(A3) Numerator and denominator are parts of one variance decomposition. **Not met, and this is measurable.** The numerator 4.380 is content-controlled, of the raw cross-half reproducibility $T_{\mathrm{cv}} = 6.902$, 2.521 (36.5%) is reproduced under within-item permutation and is subtracted. The denominator 9.402 is *not* content-controlled: 3.665 of it (39.0%) is likewise reproduced under the permutation null and is retained. The ratio therefore divides a content-controlled quantity by an uncontrolled one and is not a variance fraction of any single decomposition.

(A4) A history size n and a $1/n$ sampling-noise law, which $R(n)$ additionally needs. **Not met.** Read-out items per user range 8–181 (median 15) while the halves split unrecorded *profile rows*, so no single n indexes the estimate; and Δ_u is the output of a gradient-trained module, whose error obeys no $1/n$ law, so the Spearman–Brown aggregation used elsewhere does not apply.

What we therefore report. The quantity

$$\hat{\rho}_{\text{cross-half}} = \frac{4.380}{9.402} = 0.466 \quad [0.137,\ 0.935] \tag{23}$$

is a *descriptive cross-half reproducibility ratio*: the share of the observed between-user spread in PerFit's per-user shift that reproduces across two separately trained halves of a user's history, with the numerator content-controlled and the denominator not. It is **not** a validated fraction of true-target variance, because (A2) cannot be tested here and (A3) does not hold; and it is **not** $R(n)$, because (A4) does not hold either. Its interval is wide enough to be consistent with anything from weak to near-perfect reproducibility. It is not comparable with the R values reported for MultiPref or Community Alignment and cannot be projected to another history depth.

One piece of independent evidence bears on b_u without quantifying it: the held-out owner-versus-stranger candidate-tag margin is $+.0444\,[+.0249, +.0667]$, so the user-specific component carries behaviourally relevant information and is not purely an artifact. That supports $\theta_u \neq 0$; it does not license a variance decomposition. Settling the θ_u versus b_u split would require half-models trained under independent seeds, or read-out on disjoint item sets per half, neither available in the released artifacts, and both a retraining experiment rather than a re-analysis. Coordinate alignment, the one precondition fully satisfied, holds: both halves are read at the same intervened layers and positions of the same frozen base model, the mean half-norm ratio is 0.994, and $\cos(A_u, B_u) > 0$ for 93.6% of users.

Q.4 Why the dispersion target is not estimable here

The saved dispersion increment for PerFit is large and negative (-179.3, 95% CI $[-206.3, -146.7]$). **A negative increment is not by itself a demonstration that the estimator's independence condition fails**, and we do not present it as one: it is a difference between two levels that are not comparable, since the permutation null's $\hat{q}$ averages 16.2 against an observed 1.2, a thirteenfold inflation produced by assigning each pseudo-user observations from different real users. What we can state is descriptive and measured. Across users, $\mathrm{corr}(\hat{q}_u^A, \hat{q}_u^B) = .995$, and 99.7% of the between-user variance of $\hat{q}$ is common to the two halves, so almost nothing cross-fits away; 55% of the per-user estimates are negative although $\hat{q}$ estimates a non-negative quantity; and the mean number of users per item is 2.81, so $c_s \approx 1.8$, an order of magnitude below the $c \sim 17$ at which we argued in Section E that the cross-split leak is second order. Three features of the design are consistent with these observations, the two halves are residualised against the *same* item set and therefore share peer composition exactly, the half-models are near-duplicates of one another, and the peer counts are very thin, but **we have not isolated which of them is responsible**. The defensible conclusion is narrower than a mechanism: on this construction the quantities the dispersion estimator needs to be independent are empirically almost perfectly dependent, so the target is not estimable here, and a constructor whose halves share a collective stage, an initialisation and an evaluation item set should not be expected to support it without a redesign of the split.

R How prior evaluations map to the target-aligned evidence questions

This appendix tests the paper's motivating gap against representative recent personalized-representation evaluations. We record which scientific questions each evaluation directly targets: observation support, user binding, heterogeneity of a named target, behavioural/semantic interpretation, and downstream utility. The goal is not to argue that prior evaluations are deficient; each was designed for a different question. Instead, the comparison asks whether these evaluations already provide the target-specific inferential layer proposed here. A "partial" entry in Table 9 means that the evaluation provides relevant evidence but does not test the target-identification null defined here.

Table 9: Evidence coded in representative personalization approaches. Columns correspond to observation support, user binding, named target heterogeneity, behavioural/semantic interpretation, and downstream utility. "–" means the item is not an evaluation target of the inspected method; "partial" means related but not equivalent evidence. Full coding notes and source excerpts are included with the released audit.

Work	Personalized object	Support	Binding	Named target	Behaviour	Utility
StyleVector	user activation vector	partial	partial	partial	partial	yes
SteerX	user steering vector	–	–	–	partial	yes
Preference Heads	sparse head set	–	partial	partial	yes	yes
PerFit	collective + user shift	partial	–	partial	partial	yes
PPlug	user embedding	partial	–	–	–	yes
DEP	peer-relative latent difference	partial	partial	–	partial	yes

Across the inspected methods, user-specific information is evaluated through combinations of binding, behavioural effects, representation differences, or downstream utility. Several provide partial evidence relevant to a named target, but none implements the full sequence used by Target-Aligned Validation: prespecify $T(P_u)$, verify observation support, construct content-controlled cross-history evidence for that target, qualify dependence, and separately quantify finite-history reliability. This is the narrower evaluation gap addressed by our framework.