When Fidelity Fails to Certify Policy Selection: Decisional Anchoring for LLM User Simulators

Anonymous authors
Paper under double-blind review

Abstract

LLMs are increasingly used as user simulators in recommender systems to train agents and compare candidate policies. Existing evaluations, however, almost exclusively measure *behavioral fidelity*—how closely a simulator resembles real users—and implicitly treat such fidelity as evidence that the simulator can be trusted for downstream *policy selection*. We test this inference directly. We characterize simulators through three complementary fidelity views (probability calibration, conditional response prediction, and treatment response) and construct variants with systematically different fidelity profiles via calibration, behavioral cloning, intervention-aware training, and controlled information perturbation. Using randomized real-user logs as a supported reference for policy value, we relate each view to policy ranking, pairwise decision correctness, and decision regret. Across a large randomized recommender dataset and an anonymous bandit log we find that the views rank simulators inconsistently, that improving a targeted behavioral property does not transfer to better policy decisions, and that fidelity features yield no stable incremental predictive validity for pairwise decision correctness. Behaviorally more faithful simulators are therefore not automatically more trustworthy for policy selection. We argue that fidelity is not an intrinsic property of a simulator but a *measured relation* among the simulator, the information set, the supported policy family, the decision margin, and a sufficiently stable real-user reference, and we specify a decisional-anchoring contract for reporting it.

1 Introduction

LLMs are increasingly used to simulate user responses in recommender systems, supporting agent training, interactive evaluation, and comparison of candidate algorithms. Unlike live user experiments, they generate responses in a controlled environment, bridging offline experimentation and real-user evaluation; prior work shows that model choice, prompting, user profiles, and interaction history all materially affect simulated outcomes (Yoon et al., 2024; Zhang et al., 2023; 2025; Zhu et al., 2024).

Existing user-simulation research primarily asks whether simulated behavior matches real users—choices, preference expression, response prediction, interaction outcomes, or aggregate behavioral patterns (Yoon et al., 2024; Zhang et al., 2025; Chen, 2026). Recent work examines simulator reliability and verifiability (Zhu et al., 2024; Hsu et al., 2024; Ma et al., 2026), yet such agreement is still evaluated at the behavioral level, which we call *simulator fidelity*: it measures how well a simulator matches real users under a given task, information condition, and response format. Policy evaluation instead cares about correctness: when a simulator compares recommendation methods, can it preserve the relative ordering among policies? Once a simulator moves from *generating behavior* to *adjudicating policies*, it is the decision basis for which method is judged better.

Behavioral fidelity and decisional correctness are not equivalent by definition: a simulator better at response prediction, calibration, or a particular behavioral effect need not compare policies more accurately. Figure 1 illustrates the failure mode: a simulator can track a real user's ratings closely

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

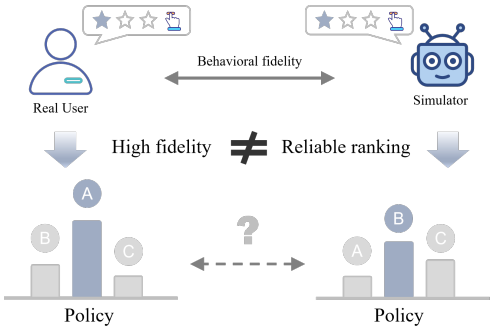


Figure 1: Behavioral similarity does not guarantee decisional agreement. Real users and a simulator may give similar ratings to the same items (top) yet induce different preference orderings (bottom): the real user prefers $A \succ B \succ C$ while the simulator prefers $B \succ A \approx C$. Small per-item rating errors suffice to reverse a pairwise comparison—the judgment a simulator makes when it selects a policy.

enough to look faithful and still order the items differently, because ranking depends on the sign of the gaps between items, not on absolute accuracy. Averaged similarity therefore does not bound the error of the comparison the simulator is asked to make, and the two diverge exactly where candidate policies are close. We therefore ask: *can behavioral fidelity of user simulators indicate the reliability of their policy comparison and selection, and do improvements in specific fidelity properties transfer to decision outcomes?* Our contributions are:

- We reframe simulator evaluation from behavioral similarity to the **relationship between fidelity and the reliability of policy selection**, separating behavioral fit from decisional validity.
- We establish a unified experimental link between multi-view fidelity measurement, controlled simulator conditions, and randomized real-user policy outcomes, testing fidelity variation under the actual use case of policy ranking and selection.
- We show that **behavioral fidelity and decisional reliability can decouple**: higher fidelity is not automatically evidence for more reliable policy selection, and the relationship is constrained by policy differences and real-user reference stability.

2 Related Work

Building LLM user simulators. LLM-based user simulation in recommendation has developed along several design paths, differing in which modules they add: Agent4Rec equips generative agents with profile, memory, and action modules (Zhang et al., 2023); the LLM-Powered User Simulator combines LLM preference logic with statistical modeling (Zhang et al., 2025); RecUserSim adds bounded-rational action and response refinement (Chen et al., 2025); PUB conditions behavior on personality (Ma et al., 2025); and UserMirrorer builds preference-aligned training data from feedback (Wei et al., 2026). This line shows how prompts, memory, persona, and task-specific training change behavior, but it typically *validates the simulator*; our target is the next inferential step—whether the metrics these choices improve certify policy selection.

Measuring realism, fidelity, and verifiability. Evaluation protocols differ in unit and reference: Yoon et al. compare item selection and preference expression against humans (Yoon et al., 2024); Lost in Simulation assesses proxy reliability for real versus simulated populations (Seshadri et al., 2026); and Chen compares simulated responses against purchase outcomes (Chen, 2026). These move from surface plausibility to outcome-anchored response validity, but their unit remains a response, population, or agent score—not the validity of a fidelity metric as a proxy for policy selection. The verifiability agenda reaches a similar boundary: Ma et al. decompose user simulation into auditable components (Ma et al., 2026) and Tau-Rec uses verifiable rewards instead of LLM judges (Narasimhan & Narasimhan, 2026); auditability establishes *how* an evaluation was produced, not that fidelity measurements predict the cost of choosing a policy from it.

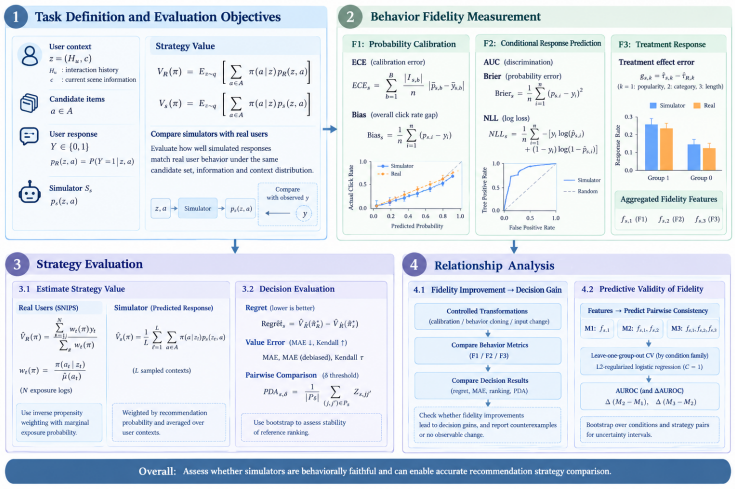


Figure 2: Overview of the decisional-anchored evaluation framework. **(1)** Task and objective: a context $z = (H_u, c)$ and candidate item a elicit a binary response Y, giving real-user and simulator probabilities $p_R(z,a)$, $p_s(z,a)$, and policy values $V_R(\pi)$, $V_s(\pi)$. **(2)** Fidelity measurement: the three views F1 (calibration), F2 (conditional response), F3 (treatment response), aggregated into $f_{s,1}, f_{s,2}, f_{s,3}$. **(3)** Strategy evaluation: real-user value by SNIPS on randomized logs, simulator value by predicted response, and the decision metrics. **(4)** Relationship analysis: controlled transformations compare behavioral metrics against decision results, and a leave-one-family-out L2 logistic regression predicts pairwise decision consistency from the fidelity features. Section 3 defines every term.

From simulation to policy validity. The closest precedent is Hsu et al., who develop counterfactually robust user models and compare simulation-based evaluation with online outcomes (Hsu et al., 2024). We likewise require a real-user reference but change the validation target: rather than asking whether a user-modeling method reproduces policy outcomes, we ask whether commonly reported fidelity measurements explain policy-selection error under controlled conditions. Randomized logs make this possible without treating observational recommendations as ground truth: KuaiRand provides randomly exposed short-video interactions (Gao et al., 2022) and OBD bandit logs with known propensities (Saito et al., 2021). Pre-LLM platforms (RecSim (Ie et al., 2019), RecSim NG (Mladenov et al., 2021)) and surveys (Stavinova et al., 2022) show that user simulation is broader than language generation, and LLM reliability work shows that leakage and prompt control can dominate apparent quality (Zhu et al., 2024). Our contribution is an external validation test for response simulators with a declared downstream policy-selection purpose; Table 3 (Appendix A) contrasts the validation unit and boundary of each line.

3 Method

Figure 2 overviews the framework. Panels 1–2 define the task, the evaluation objective, and the three behavioral fidelity views that aggregate into features $f_{s,1}, f_{s,2}, f_{s,3}$ for each simulator condition S_s; panels 3–4 define the policy-decision metrics against a randomized real-user reference and the two analyses relating fidelity to decisions.

3.1 Task Definition and Evaluation Objective

Let $z = (H_u, c)$ be the recommendation context of user u (H_u the pre-recommendation interaction history, c the current scenario information) and $a \in \mathcal{A}$ an exposed item from candidate set $\mathcal{A}$. A real user produces a binary response $Y \in \{0,1\}$ with conditional response probability $p_R(z,a) = P(Y = 1 \mid z, a)$. Each simulator condition S_s—fixed by a model, prompt, and training or calibration setting—outputs $p_s(z,a)$. Since p_R is unobservable, we evaluate simulated probabilities against log

response samples: the real event's context and exposed item are supplied to the simulator and its prediction compared with the observed response.

Let $\Pi = \{\pi_1, \ldots, \pi_J\}$ be J recommendation policies, where $\pi(a \mid z)$ is the probability that π recommends item a in context z. Policy value is the average expected feedback (expected click rate in a click task). With $q(z)$ the context distribution of evaluated users and scenarios,

$$V_R(\pi) = \mathbb{E}_{z\sim q}\left[\sum_{a\in\mathcal{A}} \pi(a \mid z)\, p_R(z,a)\right], \qquad V_s(\pi) = \mathbb{E}_{z\sim q}\left[\sum_{a\in\mathcal{A}} \pi(a \mid z)\, p_s(z,a)\right]. \tag{1}$$

The two values are the same policy's expected click rate under real users and under the simulator; for comparability the candidate set, pre-recommendation information, and context distribution are held fixed.

3.2 Behavioral Fidelity Measurement

A simulator whose probabilities match overall response frequencies may still fail to discriminate which events are more likely, and good event-level prediction need not recover real-user response differences across exposure conditions. We therefore measure fidelity from three views—probability calibration (F1), conditional response prediction (F2), and treatment response (F3)—that respectively ask whether predicted probabilities match frequencies, whether context-conditioned prediction is accurate, and whether response differences between predetermined exposure groups are reproduced; they may overlap statistically and are not assumed independent or exhaustive.

F1: Probability calibration. With evaluation sample $\{(z_i, a_i, y_i)\}_{i=1}^n$, $p_{s,i} = p_s(z_i, a_i)$, $B = 10$ equal-width bins, and $\mathcal{I}_{s,b}$ the events predicted in bin b with means $\bar{p}_{s,b}, \bar{y}_{s,b}$,

$$\mathrm{ECE}_s = \sum_{\substack{b=1 \\ |\mathcal{I}_{s,b}|>0}}^{B} \frac{|\mathcal{I}_{s,b}|}{n} \left|\bar{p}_{s,b} - \bar{y}_{s,b}\right|, \qquad \mathrm{Bias}_s = \frac{1}{n}\sum_{i=1}^{n}(p_{s,i} - y_i), \tag{2}$$

with empty bins excluded; smaller ECE is better and the sign of Bias indicates over- or underestimation.

F2: Conditional response prediction. We use the area under the ROC curve (AUC_s) for discrimination and the Brier score and negative log-likelihood (NLL) for event-level probabilistic error, clipping probabilities to $[10^{-6}, 1-10^{-6}]$. Higher AUC and lower Brier/NLL are better; since Brier and NLL are also affected by calibration, F2 and F1 are not independent.

F3: Treatment response. For a predetermined contrast set $\mathcal{K}$ (each k specifying two exposed-item groups, a feedback type, and the evaluated users), with treatment and reference event sets $\mathcal{I}_{k,1}, \mathcal{I}_{k,0}$,

$$\widehat{\tau}_{s,k} = \frac{1}{|\mathcal{I}_{k,1}|}\sum_{i\in\mathcal{I}_{k,1}} p_s(z_i,a_i) - \frac{1}{|\mathcal{I}_{k,0}|}\sum_{i\in\mathcal{I}_{k,0}} p_s(z_i,a_i), \qquad g_{s,k} = \widehat{\tau}_{s,k} - \widehat{\tau}_{R,k}. \tag{3}$$

The real contrast $\widehat{\tau}_{R,k}$ is a frozen registry value (KuaiRand popularity/category/duration clicks: 0.0924, 0.1064, -0.0149) not re-estimated on the simulated sample; resampling keeps it fixed. Smaller $|g_{s,k}|$ is better, and a direction reversal requires comparing the signs of $\widehat{\tau}_{s,k}$ and $\widehat{\tau}_{R,k}$.

Aggregated fidelity features. Each metric is standardized with the mean and sample standard deviation of a fixed 27-condition reference set, written $\mathcal{Z}$ (reused later for comparability). The three views aggregate as

$$\begin{aligned} f_{s,1} &= \tfrac{1}{2}\big[\mathcal{Z}(-\mathrm{ECE}_s) + \mathcal{Z}(-|\mathrm{Bias}_s|)\big], \\ f_{s,2} &= \tfrac{1}{3}\big[\mathcal{Z}(\mathrm{AUC}_s) + \mathcal{Z}(-\mathrm{Brier}_s) + \mathcal{Z}(-\mathrm{NLL}_s)\big], \\ f_{s,3} &= \mathcal{Z}\left(-\tfrac{1}{3}\sum_{k=1}^{3}\frac{|g_{s,k}|}{|\widehat{\tau}_{R,k}| + 10^{-6}}\right), \end{aligned} \tag{4}$$

where negated errors are averaged equally (AUC in its original direction) so larger is better, and $f_{s,3}$ normalizes each absolute error by the real contrast magnitude. The vector $\mathbf{f}_s = (f_{s,1}, f_{s,2}, f_{s,3})$ describes a condition's overall performance rather than a single user or pair; equal weighting only fixes the combination rule, and the features are deliberately not summed into one score.

3.3 Policy Decision Evaluation

Real-user value on KuaiRand is estimated by self-normalized inverse propensity scoring (SNIPS). With N exposure records, logging probability $\widehat{\mu}(a_t \mid z_t)$, and weight $w_t(\pi) = \pi(a_t \mid z_t)/\widehat{\mu}(a_t \mid z_t)$,

$$\widehat{V}_R(\pi) = \frac{\sum_{t=1}^{N} w_t(\pi)\, y_t}{\sum_{t=1}^{N} w_t(\pi)}, \qquad \widehat{V}_s(\pi) = \frac{1}{L}\sum_{\ell=1}^{L}\sum_{a\in\mathcal{A}} \pi(a \mid z_\ell)\, p_s(z_\ell, a), \tag{5}$$

where the simulator value averages predicted responses over L shared sampled contexts z_ℓ. The estimate requires non-zero logging probability for every item the target policy may recommend and a correct propensity estimate; the implementation uses an item-marginal $\widehat{\mu}(a)$ not varying with user context, checked but not validated by effective sample size and log hit rate. Real value aggregates over logged exposures while simulated value averages sampled users equally, so their difference may reflect sampling and weighting as well as response error.

Let $\widehat{\pi}_R^*$ and $\widehat{\pi}_s^*$ be the highest real-log and simulated-value policies (ties broken by a fixed order). Regret, value error, and ranking are

$$\widehat{\mathrm{Regret}}_s = \widehat{V}_R(\widehat{\pi}_R^*) - \widehat{V}_R(\widehat{\pi}_s^*), \qquad \mathrm{MAE}_s = \frac{1}{J}\sum_{j} \left|\widehat{V}_s(\pi_j) - \widehat{V}_R(\pi_j)\right|, \tag{6}$$

with Kendall's τ between the two value orderings; lower MAE and higher τ are better. We also subtract the cross-policy mean signed error and recompute MAE (de-leveled) as an error analysis, not recalibrated performance.

Pairwise comparison checks whether the simulator preserves the superiority relation of two policies. With a positive click-probability threshold δ and $j < j'$,

$$\mathcal{P}_\delta = \left\{(j,j') : 1 \le j < j' \le J,\ \left|\widehat{V}_R(\pi_j) - \widehat{V}_R(\pi_{j'})\right| > \delta\right\}, \qquad \mathrm{PDA}_{s,\delta} = \frac{1}{|\mathcal{P}_\delta|}\sum_{(j,j')\in\mathcal{P}_\delta} Z_{s,jj'}, \tag{7}$$

where $Z_{s,jj'} = 1$ if the simulated and real-log value orderings agree, 0 if they disagree, and $1/2$ for a simulated tie. Exceeding δ does not guarantee a stable reference ordering, so bootstrap checks the sign and stability of reference differences and flags reference-stable pairs by a predetermined joint rule; the flag describes uncertainty and does not remove pairs from the main analysis.

3.4 Relating Behavioral Fidelity to Policy Evaluation

We relate fidelity to decisions in two ways. The **controlled comparison** evaluates a predetermined transformation: it compares the behavioral metrics of Section 3.2 before and after, judges whether the target property improved, and compares the decision metrics of Section 3.3 under a shared setting, policy set, and reference. A target improvement with a worsened decision metric is a counterexample to the corresponding improvement relation; because one transformation may change several properties at once, this does not isolate the causal role of a single metric.

The **predictive analysis** tests whether fidelity features distinguish conditions with different decision performance, using $\mathbf{f}_s$ and $Z_{s,jj'}$ (unit (s,j,j'), $(j,j') \in \mathcal{P}_\delta$; the frozen data contain no simulated ties, so $Z_{s,jj'} \in \{0,1\}$ is a binary response). Three L2 logistic regressions ($C = 1$) add features progressively (M1: $f_{s,1}$; M2: $+f_{s,2}$; M3: $+f_{s,3}$), predicting $Z_{s,jj'} = 1$ from condition-level features only. We use leave-one-group-out validation over predetermined condition families, pooling held-out predictions and reusing the fixed standardization; performance is the AUROC of pooled held-out predictions and the increments $M2-M1$, $M3-M2$ (computed only when both classes are present). Because observations share conditions and pairs, we two-way bootstrap these units with replacement and refit within each repetition, taking the 2.5th/97.5th percentiles over jointly valid repetitions; the resampling sign proportion is a descriptive tail fraction, not a p-value. Appendix A gives the full procedure.

4 EXPERIMENTS AND RESULTS

4.1 EXPERIMENTAL SETUP

We ask three questions. **RQ1** tests whether the calibration, conditional-response, and treatment-response views judge the same conditions consistently; **RQ2** whether improving a specific fidelity property accompanies more accurate value estimation, ranking, and selection; and **RQ3** whether F1–F3 predict whether a simulator's judgments of policy superiority are reliable.

Table 2 (Appendix A) gives the datasets. KuaiRand-27K (Gao et al., 2022) is primary for all three questions: its standard log provides user history and policy construction, and its random-exposure window estimates treatment responses and real-user value. Users are split by hash into Train/Calibration/Test ($55/12.5/32.5\%$), and Test users split in time into Support history (visible history), Probe-A (fidelity), and Decision-B (real-user value), with Probe-A and Decision-B feedback excluded from training, calibration, propensity estimation, and policy construction. An artifact audit found 6,787 of 18,939 early full-Probe records belong to Probe-A and 12,152 to Decision-B; this is not training leakage, but the intended row-level separation was not achieved, so full-Probe RQ1/RQ2 results are descriptive, RQ3 and the matched baselines use Probe-A-only recomputation, and full-Probe serves as a sensitivity analysis. OBD (Saito et al., 2021) (randomized bandit logs with known propensity but no persistent user history) checks whether the RQ1/RQ2 phenomena appear with anonymous, sparse feedback; Amazon Reviews 2023 (no randomized exposure or identifiable logging policy) only supplements F1/F2 and contamination analysis. The datasets are analyzed separately without pooling; full splits and condition definitions are in Appendix A.

The primary backbone is Qwen2.5-7B-Instruct, with Qwen2.5-14B and Llama-3.1-8B for scale and family comparison; models answer a binary question given the visible history, context, exposed item, and condition-visible information, with click probability from the normalized log-probability of the answer tokens "0"/"1" (a pilot showed 99.4% hard-decoded "no click", so we use checked soft probabilities). Simulator conditions span calibration (C0 prompt-only; C1a single-token LoRA behavioral cloning; C3 Platt scaling, isotonic regression, and γ interpolation; C4/C4b intervention-aware reweighting and effect matching) and four information/training perturbations (A_{hist} history retention, B_{order} history order, D_{obs} observation masking/coarsening, E_λ intervention-aware weight); they probe whether metrics respond to clear information or training changes, not a random sample of natural simulator variation. The treatment-response registry has nine targets across popularity, category, and duration, with clicks K1, K2a, K3 feeding F3.

The KuaiRand policy set has 16 pre-registered history-based policies plus 9 added by a predetermined saturation trigger and frozen before inspection; 15 of 25 meet the support requirement, all selecting a top-100 from 7,583 candidates and recommending uniformly (Appendix A). Real value is SNIPS on 246,749 Decision-B random exposures; simulator value is computed on 1,000 stratified Test users (seed 2026); IPS, truncated SNIPS, and doubly robust estimators are sensitivity checks. The RQ3 primary analysis at $\delta = 0.005$ has 20 conditions and 59 distinguishable pairs (1,180 observations), with incremental AUROC intervals from two-way bootstraps using the same L2 fit.

4.2 RQ1: ARE BEHAVIORAL FIDELITY VIEWS CONSISTENT?

On the full-Probe descriptive scope the three backbones achieve similar but limited condition discrimination while differing markedly in probability level and treatment response. Qwen2.5-7B has AUC 0.594 and ECE 0.214; Qwen2.5-14B raises AUC to 0.626 but worsens ECE to 0.350; Llama-3.1-8B has AUC 0.624, ECE 0.369, overestimating clicks by about 0.37—similar F2 does not imply similar F1. Intervention response again diverges: for K1 every backbone's high–low popularity contrast is too small (deviations -0.055, -0.027, -0.072 from the real $+0.092$); scaling Qwen 7B$\rightarrow$14B shrinks the K1 gap but does not fix its sign; for the category contrast both Qwen models overestimate while Llama nearly matches; and for K3 the real duration effect is negative, yet all three uncalibrated simulators produce a positive effect. Controlled information changes are targeted but uneven: removing all history drops AUC $0.594 \rightarrow 0.515$ and nearly eliminates output variance; masking category shrinks the category-effect gap from $+0.134$ to $+0.019$; and providing popularity quartiles pushes the K1 gap across zero to $+0.078$.

RQ1 is answered negatively: F1, F2, and F3 do not yield interchangeable orderings (Probe-A-only and full-Probe aggregates correlate at $0.996/0.997/0.993$, so the segmentation correction does not

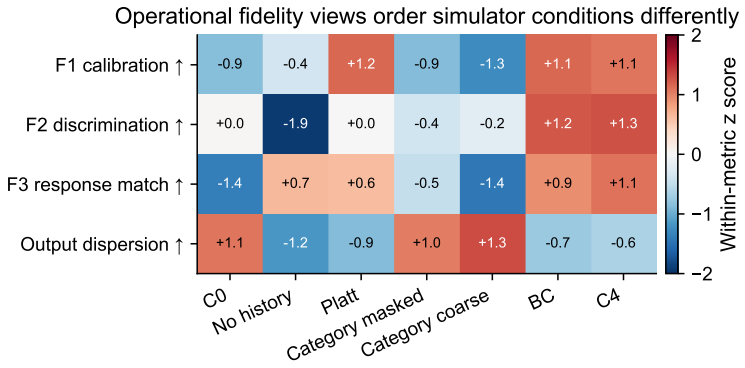


Figure 3: Representative metrics of the three fidelity views, plus output dispersion, rank conditions differently (RQ1). Rows are oriented so higher is more faithful (negative ECE, AUC, registered K1–K3 aggregate); dispersion is the standard deviation of predicted click probability. Each row is standardized within itself across the seven displayed conditions, so zero is that row's mean over them and colors and values support within-row comparison only.

change this), and we do not claim the views are orthogonal, exhaustive, or replaceable by a scalar total. Figure 3 makes the decoupling visible: no condition is favorable on every row. Removing history raises the F3 row while producing the worst F2; full Platt leads calibration but sits at the row mean on F2; the behaviorally trained conditions lead F2 and F3 yet fall below the mean on output dispersion, where the widest-dispersion conditions (C0, category-coarsened) are in turn the weakest on F3.

4.3 RQ2: Does Improving Fidelity Transfer to Better Policy Decisions?

On the full-Probe descriptive scope, marginal calibration does not automatically transfer to value accuracy on the policy-induced support. In the C3 gamma sweep, full Platt calibration reduces ECE monotonically while AUC stays at about 0.594, yet policy-value MAE nearly doubles over the same sweep (Figure 4a): calibration pulls probabilities toward the overall click rate of about 0.17, whereas the reference value of the supported policy's item set is near 0.35. This is a support-mismatch counterexample, not evidence that calibration and decision value are generally unrelated. Behavioral cloning gives a different profile: C1a reduces ECE to 0.006, raises AUC to 0.643, and improves the duration-effect gap when duration is explicit, yet retains only about 1% of C0's within-user output variance and 9% of its between-user variance (Figure 4b), a compression we call *responsiveness collapse*. The Platt and C4 bars in the same panel show the compression is not specific to cloning: each keeps about 2% of C0's within-user variance, while their between-user variance differs (3% and 13% of C0). Relative to C1a, the intervention-aware C4 sweep improves some F3 gaps (best combined F3 at $\lambda = 2$) but the Pareto surface stays flat on the coarse-margin decision task, and C4b's calibration side effect collapses under effect matching—an unresolved calibration/Test asymmetry.

Matched statistical baselines. Matched statistical response models give a stronger counterexample. On Probe-A, four logistic and LightGBM variants have standardized F1, F2, and F3 all above C0, yet regret 0.0423, Kendall's τ about -0.58 to -0.62, and PDA only 0.017–0.051, whereas C0 has regret 0.0019, $\tau = 0.600$, PDA 0.966 (Table 4, Appendix B); relative to C0 and C1a they are worse on regret, τ, PDA, and de-leveled value error (all Holm-corrected $p \leq 0.0224$); their F1 paired resampling is incomplete, so only point estimates are reported.

Local repair versus global decision. On independent validation users, $\gamma = 0.50, 0.75, 1.00$ and C1a/C4/C4b all repair the pre-locked `knn_nb50` vs `knn_mix075` pair ($\gamma = 0.25$ does not), yet globally the behaviorally trained conditions still select `covis_min20` with regret 0.0105 above C0's 0.0019: repairing one locked pair does not recover the whole ranking, and the pre-registered hypothesis that C4 reduces regret more than marginal calibration is unsupported.

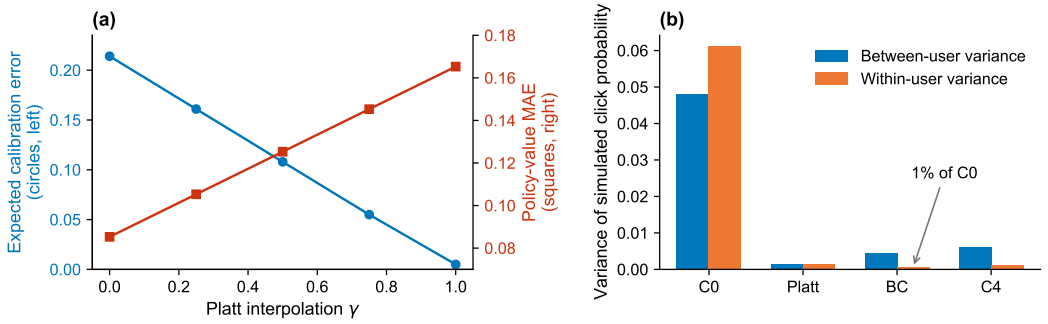


Figure 4: Better marginal calibration comes with worse policy-value accuracy and compressed responsiveness (RQ2). **(a)** C3 gamma sweep on two axes: as the Platt interpolation γ rises from 0 to 1, expected calibration error falls monotonically $0.214 \rightarrow 0.005$ (left axis, circles) while policy-value MAE rises $0.085 \rightarrow 0.165$ (right axis, squares); AUC is unchanged at 0.594 throughout and is not plotted. **(b)** Between-user and within-user variance of simulated click probability for four conditions. Relative to C0, full Platt ($\gamma = 1$) and C4 keep about 2% of the within-user variance and C1a behavioral cloning (BC) about 1%, the responsiveness collapse discussed in the text.

Table 1: Decision effectiveness on the post-saturation 25-policy extension as the margin narrows. Panel A reports condition×pair crossing rows, not independent samples. Panel B reports the revised Probe-A family leave-one-out AUROC at $\delta = 0.005$; at $\delta = 0.02$ AUROC is undefined (single class).

Panel A: decision rows

Margin δ	Distinguishable pairs	Crossing rows	Disagreements	Rate
0.020	14/105	280	0	0.000
0.010	34/105	680	6	0.009
0.005	59/105	1,180	75	0.064

Panel B: revised held-out AUROC at $\delta = 0.005$

M1: F1	M2: F1+F2	M3: F1+F2+F3
0.409	0.463	0.412

4.4 RQ3: Can Fidelity Predict the Reliability of Policy Selection?

Coarse-margin saturation. On the original 16-policy surface seven policies meet the support criterion; at $\delta = 0.02$ every LLM condition selects `knn_click` (regret 0.0008, PDA 1.0, Kendall ≈ 0.714) and all 162 condition–pair observations of the pre-registered E5 analysis are correct, so the F2/F3 incremental hypotheses are unidentifiable here (single-class outcome).

Extended policy surface and decision margin. After the registered saturation trigger, the independently frozen 25-policy extension raises the supported set to 15 and changes the real optimum to `knn_mix075` ($V_R = 0.3492$); as the margin narrows, the disagreement rate grows from zero at $\delta = 0.02$ to $6/680$ at $\delta = 0.01$ and $75/1{,}180$ at $\delta = 0.005$ (Table 1A).

Reference stability. All 75 disagreements occur on the 28 reference-unstable pairs; the 31 stable pairs have 620 crossing rows with no disagreement and a single-class outcome that leaves AUROC undefined (Figure 5). Disagreements therefore cannot be separated from real-user ranking uncertainty.

Criterion predictive validity. After removing the F1 slope term and unifying the fit, the Probe-A held-out AUROCs are 0.409 (M1), 0.463 (M2), 0.412 (M3): adding F2 gives a positive and adding F3 a negative point estimate, but every interval spans zero (Figure 6) and the signs are unstable—among 88 influence-point deletions $M2 - M1$ reverses 6 times and $M3 - M2$ 3 times,

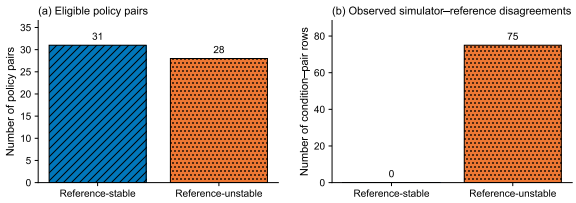


Figure 5: Observed disagreements occur only where the real-user ordering is unstable (RQ3, $\delta = 0.005$). **(a)** Eligible policy pairs split by the frozen joint stability rule: 31 reference-stable, 28 reference-unstable. **(b)** Condition–pair rows on which a simulator contradicts the reference ordering: 0 in the stable group (of 620 crossing rows) and all 75 in the unstable group.

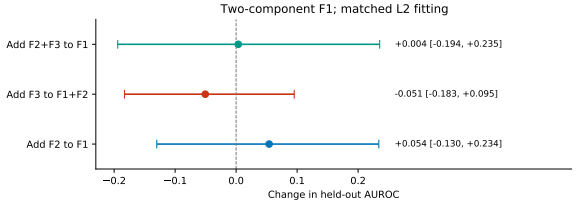


Figure 6: No nested feature addition yields a stable AUROC increment (RQ3). Points are differences of family leave-one-out held-out predictions at $\delta = 0.005$ under the two-term F1 and matched L2 fitting, with 95% bootstrap percentile intervals from the 9,976 jointly valid of 10,000 two-way repetitions. "Add F2 to F1" is $M2 - M1$ and "Add F3 to F1+F2" is $M3 - M2$ (primary); "Add F2+F3 to F1" is the supplementary $M3 - M1$. The dashed line marks zero, which every interval crosses.

and the full-Probe sensitivity mirrors this (Appendix B). Disagreements concentrate on interpretable conditions but inherit OPE uncertainty—observed differences, not a general mechanism. The results therefore establish neither stable incremental validity nor the absence of predictive information.

5 Conclusion

This study fails to identify a fidelity view that reliably certifies policy selection. The three operational views are not interchangeable, and proxy improvements can fail at the intended decision: calibration improves ECE yet worsens value accuracy, behavioral cloning compresses within-user variation, and matched statistical baselines beat C0 on all three composites yet rank policies poorly. On the post-saturation fine-margin extension, both predictive increments span zero and reverse under influence-point deletion, and every observed disagreement falls on an OPE-unstable pair. On the coarse-margin surface the tested simulators expose no consequential pairwise error, a bounded non-failure result that the same saturation prevents from becoming a screening endorsement. Fidelity is therefore not an intrinsic adjective but a *measured relation* among simulator, information set, supported policy family, decision margin, and a sufficiently stable real-user reference; decisional-anchored evaluation should declare these elements and validate regret or pairwise correctness where the reference supports the intended discrimination.

Limitations. The F1 definition and predictive analysis were revised after seeing results (not pre-registered), and the two-term F1 does not cover all calibration properties. The stable reference subset contains no errors usable for estimating discrimination, so the null above is fragility rather than support. The early artifact shares evaluation rows between fidelity and simulator value (not training leakage). Cross-domain evidence gives directional decoupling, not universal effect sizes; the policy family is restricted to supported top-100 policies, the comparison set is incomplete (C1b, C2, F4 not executed), and because a simulator qualifying at $\delta = 0.02$ may fail at $\delta = 0.005$, evaluations should publish decision-effectiveness curves over chosen margins. We do not evaluate subgroup fidelity, fairness, provider exposure, or long-term welfare; simulators cannot replace user studies or human oversight (Appendix B).

Reproducibility Statement

All datasets are public: KuaiRand-27K (Gao et al., 2022), OBD (Saito et al., 2021), and Amazon Reviews 2023. Sections 3.2–3.4 give complete metric definitions and Appendix A gives splits, simulator conditions, policy construction, fitting parameters, and the bootstrap procedure. Metrics and analyses are implemented in a frozen server release; aggregated result tables, figure-generation code, and a checksum manifest form the shareable reproduction layer. Prediction-level or other derived artifacts are released subject to dataset licenses and derived-redistribution terms; where redistribution is not permitted we provide acquisition instructions, schemas, environment information, and scripts to reconstruct results from legally obtained source data.

Ethics Statement

This study performs secondary analysis of existing research datasets and does not recruit participants. It does not establish subgroup fairness, provider-side exposure effects, long-term welfare, or ecosystem safety. Matched-context simulators must not replace user studies, online evaluation, or human oversight in decisions that affect people. Artifact release is governed by verified dataset terms and applicable institutional requirements.

AI Use Statement

In this work we used generative AI tools for language editing and for organizing project materials during manuscript preparation. We did not use generative AI tools to generate research ideas, to design experiments, or to fabricate or alter data, results, or citations; those tasks are not applicable to the use of AI here. We have reviewed all AI-assisted work, and the authors verified every claim, citation, analysis, and statement against the frozen experimental artifacts. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

References

L. Chen. Simulated customers never walk away: Decision fidelity of LLM user simulators measured against real purchase outcomes, 2026. arXiv:2606.20708. https://doi.org/10.48550/arXiv.2606.20708.

L. Chen, Q. Dai, Z. Zhang, X. Feng, M. Zhang, P. Tang, X. Chen, Y. Zhu, and Z. Dong. RecUserSim: A realistic and diverse user simulator for evaluating conversational recommender systems, 2025. arXiv:2507.22897.

C. Gao et al. KuaiRand: An unbiased sequential recommendation dataset with randomly exposed videos. In *CIKM 2022*, pp. 3953–3957, 2022. https://kuairand.com/.

C.-W. Hsu et al. Minimizing live experiments in recommender systems: User simulation to evaluate preference elicitation policies. In *SIGIR 2024*, pp. 2925–2929, 2024.

E. Ie, C.-W. Hsu, M. Mladenov, V. Jain, S. Narvekar, J. Wang, R. Wu, and C. Boutilier. RecSim: A configurable simulation platform for recommender systems, 2019. arXiv:1909.04847.

C. Ma, X. Wanyan, D. Hettiachchi, Z. Xu, Y. Ren, and J. Chan. Verifiable user simulation for search and recommendation systems. In *SIGIR 2026*, 2026. https://doi.org/10.1145/3805712.3808645.

C. Ma et al. PUB: An LLM-enhanced personality-driven user behaviour simulator for recommender system evaluation. In *SIGIR 2025*, 2025. https://doi.org/10.1145/3726302.3730238.

M. Mladenov, C.-W. Hsu, V. Jain, E. Ie, C. Colby, N. Mayoraz, H. Pham, D. Tran, I. Vendrov, and C. Boutilier. RecSim NG: Toward principled uncertainty modeling for recommender ecosystems, 2021. arXiv:2103.08057.

B. S. Narasimhan and K. R. Narasimhan. Tau-Rec: A verifiable benchmark for agentic recommender systems, 2026. arXiv:2606.10156.

Y. Saito, S. Aihara, M. Matsutani, and Y. Narita. Open bandit dataset and pipeline: Towards realistic and reproducible off-policy evaluation. In *NeurIPS Datasets and Benchmarks*, 2021. arXiv:2008.07146.

P. Seshadri, S. Cahyawijaya, A. Odumakinde, S. Singh, and S. Goldfarb-Tarrant. Lost in simulation: LLM-simulated users are unreliable proxies for human users in agentic evaluations, 2026. arXiv:2601.17087.

E. Stavinova, A. Grigorievskiy, A. Volodkevich, P. Chunaev, K. Bochenina, and D. Bugaychenko. Synthetic data-based simulators for recommender systems: A survey, 2022. arXiv:2206.11338.

T. Wei, H. Guo, Y. Du, Z. Sun, H. Chen, D. Wang, and J. Zhang. Mirroring users: Towards building preference-aligned user simulator with user feedback in recommendation. In *ACL 2026*, pp. 6046–6071, 2026. https://doi.org/10.18653/v1/2026.acl-long.273.

S.-e. Yoon, Z. He, J. M. Echterhoff, and J. McAuley. Evaluating large language models as generative user simulators for conversational recommendation. In *NAACL 2024*, pp. 1490–1504, 2024. https://doi.org/10.18653/v1/2024.naacl-long.83.

A. Zhang, Y. Chen, L. Sheng, X. Wang, and T.-S. Chua. On generative agents in recommendation, 2023. arXiv:2310.10108.

Z. Zhang, S. Liu, Z. Liu, R. Zhong, Q. Cai, X. Zhao, C. Zhang, Q. Liu, and P. Jiang. LLM-powered user simulator for recommender system. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(12):13339–13347, 2025. https://doi.org/10.1609/aaai.v39i12.33456.

L. Zhu, X. Huang, and J. Sang. How reliable is your simulator? analysis on the limitations of current LLM-based user simulators for conversational recommendation, 2024. arXiv:2403.16416.

A Extended Experimental Setup

Table 2: Dataset statistics and their role in each research question. "Main" is full estimation and testing; "Supp." checks a specific phenomenon only.

Dataset	Evaluation sample used	RQ1	RQ2	RQ3
KuaiRand-27K	Probe-A: 6,787 events, 4,632 users; Decision-B: 246,749 events, 8,561 users; V_S: 1,000 users	Main	Main	Main
OBD	Probe-A: 40,000 rows; Decision-B: 1,500 contexts	Supp.	Supp.	–
Amazon Reviews 2023	8,655 probes; 3 prompts $\times$ 3 backbones	Supp.	–	–

User splits and probe construction. KuaiRand-27K[1] contains 322,278,385 standard log records, 1,186,059 random-exposure records, 27,285 random-window users, and 7,583 candidate items. Users partition deterministically by hash into Train (14,969), Calibration (3,508), and Test (8,808) with no overlap. Test users split in time into Support history, Probe-A (6,787 events, 4,632 users), and Decision-B (246,749 events, 8,561 users); Probe-A and Decision-B feedback never enters training, calibration, propensity estimation, or policy construction. Simulator value V_S is computed on 1,000 stratified Test users (seed 2026), each user–item pair inferred once and shared across policies.

Simulator conditions. Each condition fixes a model, prompt, and training/calibration configuration: C0 prompt-only; C1a single-token LoRA behavioral cloning on 12,000 Train samples; C3 Platt scaling, isotonic regression, and γ-interpolation between C0 and full Platt; C4 behavioral-cloning samples reweighted by predetermined extreme-treatment metrics; C4b C4 plus approximate effect matching; and perturbations A_{hist}, B_{order}, D_{obs}, E_λ as defined in the main text. The treatment-response registry has nine targets across popularity, category, and duration, with K1/K2a/K3 clicks feeding F3. RQ3 uses standardized F1–F3 from the frozen 27-condition reference set, with direction transforms and standardization parameters fixed before the family leave-one-out validation.

Policy construction and support. The registry has 25 policies: 16 history-based policies (popularity–personalization blends, popularity debiasing, nearest-neighbor, co-visitation) form the pre-registered set, and 9 are added by a predetermined saturation trigger and frozen before inspection. Fifteen meet the support requirement (effective sample size $\geq$ 300, log hit rate $\geq$ 0.5%); all select a top-100 from 7,583 candidates and recommend uniformly, using only Train/Calibration/Support data. Real value is SNIPS on Decision-B random exposures (log probabilities from Train/Calibration random-exposure frequency via 0.5 additive smoothing); IPS, truncated SNIPS, and doubly robust estimators are sensitivity checks.

Fitting, bootstrapping, and validity. RQ3 fits three nested L2 logistic regressions with `LogisticRegression(penalty="l2", C=1.0, solver="lbfgs", tol=1e-4, max_iter=2000, random_state=2026)`, increasing `max_iter` to 20,000 once on non-convergence; point estimate, two-way bootstrap, and influence-point deletion share this procedure. The primary margin is $\delta = 0.005$ with 20 conditions and 59 distinguishable pairs (1,180 observations); intervals use 10,000 condition$\times$pair two-way bootstraps with the product of resample counts as the row weight and family leave-one-out refit within each repetition. Twenty-four repetitions are invalid (single-class fold); intervals use the 9,976 jointly valid repetitions, and the resampling sign proportion is a descriptive tail fraction, not a null-calibrated p-value. Reference stability uses the frozen joint rule (sign probability ≥ 0.95, margin probability ≥ 0.90).

Standardization and aggregation details. Each raw metric is transformed by the direction rules of Eq. 4 and standardized using the mean and sample standard deviation of the fixed 27-condition reference set, with $\text{ddof} = 1$ and 10^{-12} added to the denominator to avoid division by zero; the same reference statistics are reused in all later evaluations. The F1 aggregation uses two terms (ECE, absolute bias), the F2 aggregation three terms (AUC, Brier, NLL), and the F3 aggregation the normalized popularity, category, and duration contrasts; ECE, Bias, and the individual $g_{s,k}$ are reported alongside the composites to retain their specific meaning.

[1] `https://kuairand.com/`

Related-work positioning. Table 3 contrasts the validation unit, reference, and boundary of each line of related work.

Table 3: Validation unit, reference, and boundary relative to this study for each line of related work.

Line	Validation unit / reference	Boundary relative to this study
Configurable simulator platforms	Trajectory and dynamics / specified generative process	Broader interaction models; different criteria
LLM behavioral evaluation	Response, population, or agent outcome / observed behavior	Fidelity metrics usually not validated against policy regret
Sim-to-real policy validation	Policy outcome of a specific user model / online outcome	Closest precedent; different validation target
This study	Fidelity metrics across conditions / randomized-log policy ranking under support constraints	Bounded metric-to-decision criterion validity; matched binary responses, short-horizon policies

B Additional Results and Sensitivity Analyses

Fidelity atlas standardization. Figure 3 (Section 4.2) is standardized within each row across the seven displayed conditions, which spreads the contrast for reading. The F1–F3 features used in RQ3 are instead standardized on the fixed 27-condition reference set (Section 3.2), so the cell values support within-row comparison and are not the feature values entering the predictive analysis.

Matched baseline table. Table 4 reports the matched statistical baselines. All four have F1, F2, and F3 point estimates above C0 yet markedly worse policy ranking and selection (regret 0.0423, $\tau \approx -0.58$ to -0.62, PDA 0.017–0.051), against C0's regret 0.0019, $\tau = 0.600$, and PDA 0.966.

Table 4: Matched statistical baselines have F1, F2, F3 point estimates above C0 but worse policy ranking and selection. F1 is the two-term standardized average of ECE and absolute bias; composites do not imply every component improves.

Model	F1	F2	F3	Regret	Kendall's τ	PDA	De-leveled MAE
C0 LLM	-0.140	-0.350	-1.284	0.0019	0.600	0.966	0.0033
C1a behavior cloning	0.894	1.133	1.168	0.0105	0.600	0.864	0.0070
Logistic, raw	0.875	1.239	2.202	0.0423	-0.619	0.051	0.0094
Logistic, Platt	0.906	1.239	2.191	0.0423	-0.581	0.051	0.0093
LightGBM, raw	0.801	1.354	2.319	0.0423	-0.600	0.017	0.0087
LightGBM, Platt	0.830	1.355	2.294	0.0423	-0.600	0.017	0.0086

Full-Probe sensitivity. The full-Probe sensitivity of the predictive-validity models gives AUROCs 0.481 (M1), 0.508 (M2), and 0.447 (M3), with increments $M2 - M1 = +0.027\ [-0.118, 0.243]$ and $M3 - M2 = -0.061\ [-0.200, 0.087]$, and 3 and 2 sign reversals under influence-point deletion—mirroring the Probe-A conclusions. Figure 6 (Section 4.4) shows the Probe-A incremental AUROC and bootstrap intervals.

Cross-domain stress tests. Within KuaiRand, Qwen2.5-14B improves some condition metrics and the K1 gap while Llama nearly matches the category intervention but misses the probability level; none produces a pairwise error above $\delta = 0.02$ on the coarse-margin supported surface, a directional model-family stress test rather than evidence of monotone scale improvement. On OBD the uncalibrated Llama condition has ECE 0.790 and mean predicted reward 0.794 (real base rate ≈ 0.0034) with $F3^* = -0.968$, yet selects the OPE reference optimum `soft_T0.5` with zero regret, whereas C1a and C4 (ECE 0.0011/0.0015) each select `pop_a1_pos1` with regret 0.000289 and PDA 0.286/0.357; on five conditions the F3–decision Spearman association is $r_s = -0.95$ ($p = 0.05$), treated as descriptive. On Amazon the pre-registered threshold is not reached: for Qwen2.5-7B the real-minus-anonymous AUC difference is $+0.0245\ [0.0137, 0.0352]$ against a required lower bound above 0.03, Qwen2.5-14B is smaller and non-significant, and the Llama anonymous arm

degenerates to constant probability 1.0 (AUC 0.5); under the registered criterion in-prompt statistics thus explain the measured performance, which does not show contamination is absent.

Iterated fit correction control. Holding sampling and fitting fixed and changing only the F1 definition isolates the net effect of removing the slope term: point estimates shift $M1 = 0.4498 \rightarrow 0.4085$, $M2 = 0.5260 \rightarrow 0.4626$, $M3 = 0.4249 \rightarrow 0.4121$. Changing only the fitting procedure relative to the archived unpenalized implementation leaves the point estimates unchanged; the difference appears only in bootstrap intervals and the deletion analysis. The revision thus affects point estimates mainly through the F1 definition, not the fitting procedure.

Baseline F1 comparability. Under the revised two-term F1, the matched baselines' ensemble F1 values are logreg_raw $+0.8749$, logreg_platt $+0.9060$, lgbm_raw $+0.8009$, lgbm_platt $+0.8300$ (from three-term $+1.1049$, $+1.1256$, $+1.1978$, $+1.2172$). Of the eight within-scope F1 comparisons after Holm correction, four are significant (logreg_raw vs C0, logreg_platt vs C0, lgbm_raw vs C0, lgbm_platt vs C1a); against C0 all four point differences remain positive, but the number of Holm-significant members shrinks relative to the three-term definition. Because the paired resampling has missing repetitions and incomplete corrections, only point estimates are reported.

OBD direction table and excluded results. Under the revised F1 the OBD fidelity–decision direction table is recomputed with each dataset's own setting and the Spearman direction is unchanged from the three-term version. Old soft-label prediction results are not updated in step with the two-term F1 and are excluded from this round's evidence; the full-Probe version of the early unpenalized procedure is covered by the plan exemption, and Amazon does not use F1. All results use the corrected frozen-server implementation.

Intervention-registry alignment. The treatment-response registry requires task alignment: K1 directly represents popularity weighting while K2/K3 only partly represent category/affinity and duration mechanisms, and no registered target characterizes neighborhood aggregation, co-visitation, conversational structure, novelty, creator exposure, or long-term satisfaction. The pre-registered incremental claim that F3 improves policy-correctness prediction beyond F1/F2 therefore concerns only K1–K3, and this revision round provides no stable support; the results are not evidence against intervention fidelity in general, and any extended basis must be specified before its outcomes are inspected.