ADAPTATION OR INFLUENCE? AUDITING SELF-UPDATING PERSONALIZED AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-updating personalized agents learn from feedback generated through their own interactions with users. Improved feedback may therefore arise from *Adaptation*, where the agent better learns a user's existing preferences, or from *Influence*, where the agent's previous actions change those preferences over time. Distinguishing Adaptation from Influence is difficult because both can produce better observed feedback. We study how to separate these effects in self-updating personalization loops. We introduce a controlled preference check that queries a preference dimension on a schedule fixed independently of the current personalized action. We further decompose action-caused preference change into a fixed-sequence effect along an already realized action sequence, a complete-loop effect when the adaptive interaction is rerun, and their difference, the additional-loop effect, which measures the additional change observed when subsequent feedback, profile updates, and actions are allowed to adapt. Across three simulated user-response mechanisms, two personalization systems, and five influence levels, ordinary feedback tracks hidden preference change poorly, with a macro Spearman correlation of 0.133. In contrast, the controlled check achieves a macro correlation of 1.000. However, simply using the preset influence strength also achieves 1.000, meaning that higher influence settings already tend to produce more hidden preference change. We therefore consider a harder comparison between the two systems at the same influence level. Here, the controlled check reaches 66.0% balanced accuracy in identifying the system with less hidden preference change, compared with 50.0% for always selecting the fixed-profile system. We also test whether self-updating creates additional Influence beyond the effects of the actions already taken. In the prespecified delayed-response setting, this additional change is only 0.00295, below our 0.01 threshold for practically meaningful amplification. These results show that, in our controlled simulation, scheduled preference queries can help distinguish Influence from Adaptation when ordinary feedback is ambiguous, while behavioral divergence does not necessarily imply substantial recursive amplification. Our findings establish a controlled synthetic benchmark and auditing framework, rather than a calibrated measure of real-user harm or welfare.

1 INTRODUCTION

Personalized agents increasingly update persistent user profiles from their own interactions with users. This creates a feedback loop: the agent chooses what a user sees, observes the user's response, updates its representation of the user, and acts again. In such a loop, improved feedback can arise through two fundamentally different mechanisms. In *Adaptation*, the user's underlying preferences remain relatively stable and the agent learns them better. In *Influence*, the agent's previous actions change the user's preferences over time, so later actions become better aligned with these changed preferences. Both mechanisms can produce better observed feedback, making them difficult to distinguish.

Consider a movie recommender that repeatedly recommends one genre and subsequently receives higher ratings for that genre. The system may have discovered an existing preference, or its repeated recommendations may have shifted the user's preference toward that genre. Ratings alone do not reveal which occurred. More generally, when an agent learns from feedback that its own past actions helped produce, observed improvement is no longer sufficient evidence of better adaptation.

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.

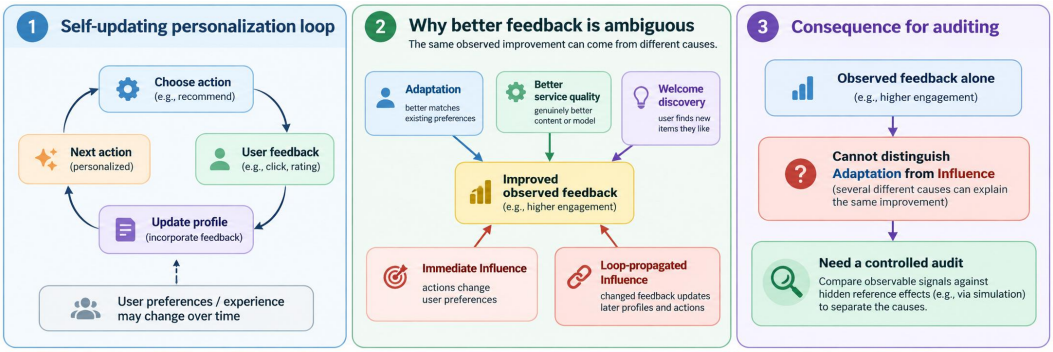


Figure 1: **Improved feedback is ambiguous in self-updating personalization.** The agent repeatedly acts, observes feedback, and updates its user profile, while its actions may also change the user's preferences over time. Improved feedback can therefore reflect *Adaptation* to existing preferences, better service, direct *Influence* on user preferences (including changes the user may welcome) or Influence propagated through later feedback and profile updates. Observed feedback alone cannot distinguish these causes, motivating a controlled audit.

This issue becomes increasingly important as personalized agents maintain persistent user representations and self-updating systems automatically revise profiles, skills, or behavior from interaction histories (Salemi et al., 2024; Liang et al., 2026; Zhang et al., 2026; Tao et al., 2026; Li et al., 2026; Wang et al., 2026; Azhar et al., 2026; Tang et al., 2026). Existing work primarily studies how to improve personalization from user feedback. We instead ask how to determine whether that feedback reflects *Adaptation* to an existing user or *Influence* on the user being modeled.

We study this question in a controlled simulation where preference change can be observed through paired counterfactual worlds. Two challenges arise. First, ordinary feedback mixes service quality with preference change and therefore need not reliably indicate Influence. We introduce a *controlled preference check*: a scheduled query of a preference dimension chosen independently of the current personalized action. The check provides an observable signal that can be compared against hidden preference change without directly accessing the simulator state.

Second, Influence can arise at different stages of the feedback loop. Actions already taken may directly change user preferences, while those changed preferences can also alter later feedback, profile updates, and subsequent actions. We therefore distinguish three quantities. The *fixed-sequence effect* measures action-caused change along an already realized sequence of actions and profiles. The *complete-loop effect* reruns the adaptive interaction so that changed feedback can affect later profiles and actions. Their difference, the *additional-loop effect*, measures the extra change introduced through this feedback-driven adaptation. This decomposition separates direct Influence from Influence propagated through the self-updating loop.

We evaluate these ideas using MovieLens-1M-grounded simulated users, three user-response mechanisms, fixed- and updating-profile systems, and five influence levels. Across six response-mechanism–system strata, the controlled preference check achieves a macro Spearman correlation of 1.000 with hidden complete-loop benchmark-defined change, compared with 0.133 for ordinary feedback. Because configured influence level is itself strongly monotonic with hidden change at this cross-dose level, we additionally examine a harder same-dose diagnostic: the controlled check reaches 66.0% balanced accuracy for distinguishing the lower-change system, versus 50.0% for always selecting the fixed system. At the same time, the decomposition rejects a stronger recursive-amplification claim: in the prespecified delayed-response condition, later feedback and profile updates add only 0.00295 additional change, below our frozen 0.01 practical threshold. Thus, changed behavior within the loop does not necessarily imply substantial additional Influence.

Our study is diagnostic rather than normative. Preference change is not inherently harmful: Influence may represent welcome discovery or unwanted change. We therefore evaluate against a fixed synthetic definition of benchmark-defined unwanted change and do not interpret the resulting quantities as calibrated measures of consent, welfare, or real-user harm. Our contributions are:

1. We formulate the central ambiguity in self-updating personalization as the distinction between *Adaptation*, where an agent better learns existing user preferences, and *Influence*, where the agent's actions change those preferences.
2. We introduce a controlled preference check for detecting preference change from observable interactions, and show that it recovers the ordering of hidden complete-loop benchmark-defined Influence substantially more reliably than ordinary feedback in our controlled simulation, while a harder same-dose analysis exposes the limits of this aggregate result.
3. We decompose Influence into fixed-sequence, complete-loop, and additional-loop effects, separating change along actions already taken from additional change observed when subsequent feedback, profile updates, and actions are allowed to adapt.
4. We show an important boundary of the feedback-loop hypothesis: although influenced and no-influence trajectories can diverge substantially, repeated profile updating does not produce practically meaningful additional Influence in the prespecified delayed-response setting.

2 Related Work

Personalized and self-updating agents. Personalized language models and persistent agents adapt generation or actions to individual users (Salemi et al., 2024; Liang et al., 2026; Zhang et al., 2026; Tao et al., 2026), while recent recommender and agent systems increasingly automate profile, skill, model, or deployment updates (Li et al., 2026; Wang et al., 2026; Azhar et al., 2026; Tang et al., 2026). These works primarily study how to improve personalization or self-update effectively. We instead study a consequence of repeated adaptation: whether improved feedback reflects *Adaptation* to existing user preferences or *Influence* that changes those preferences.

Preference influence and performative feedback. Performative prediction studies settings in which deployed models change the data they later observe (Perdomo et al., 2020). Related work on auto-induced distribution shift, preference influence, and user tampering shows that optimizing observed responses can alter future targets or users (Krueger et al., 2020; Carroll et al., 2022; Kasirzadeh & Evans, 2023). Because preference change is not necessarily unwanted, meta-preferences are also important (Ashton & Franklin, 2022). Our focus is complementary: we use controlled paired worlds to distinguish Adaptation from Influence and to separate change caused along an existing action sequence from additional change propagated through later feedback and profile updates.

Longitudinal evaluation. Longitudinal benchmarks intervene at different points in the history of persistent user-conditioned agents and trace downstream failures (Qian et al., 2025). Our study instead asks whether observable interactions can recover hidden preference change without access to latent user state, and decomposes that change into fixed-sequence, complete-loop, and additional-loop effects. This makes it possible to test both whether ordinary feedback is a reliable signal of Influence and whether self-updating feedback loops add further Influence beyond the actions already taken.

3 What Should Be Measured?

The distinction between *Adaptation* and *Influence* is causal: Adaptation changes the agent's representation of a relatively stable user, whereas Influence occurs when the agent's actions change the user's underlying preferences. Once such action-caused change occurs, a second question arises in a self-updating loop: how much change is attributable to the actions already realized, and how much additional change appears because altered feedback changes later profiles and actions?

We answer this question with three operational effects. The *fixed-sequence effect* measures action-caused change along one realized sequence of actions and profile versions by replaying that same sequence with and without action-caused Influence. The *complete-loop effect* instead reruns both adaptive worlds, allowing changed feedback to modify later profiles and actions. Their difference is the *additional-loop effect*:

$$\Delta_{\text{add}} = \Delta_{\text{loop}} - \Delta_{\text{fixed}}. \tag{1}$$

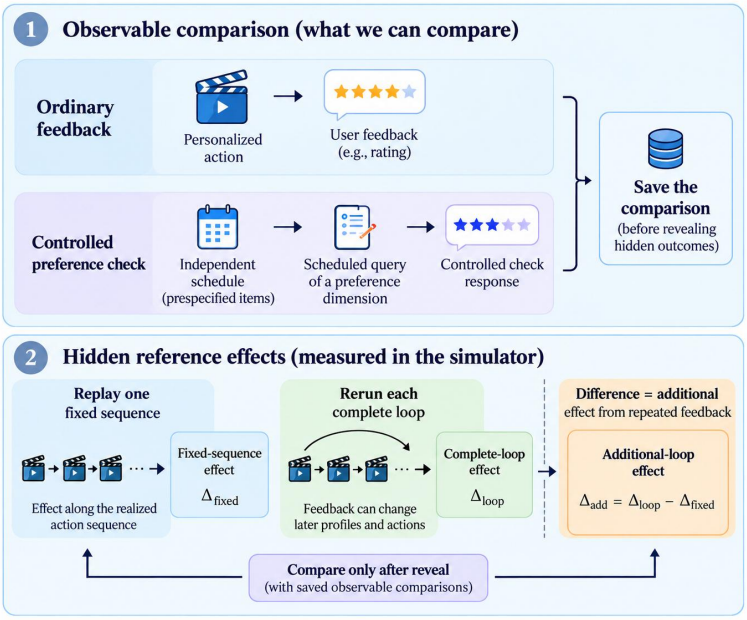


Figure 2: **Framework for auditing self-updating personalization loops.** The observable side compares ordinary feedback with a controlled preference check whose queried dimension follows an independently fixed schedule, and saves this comparison before hidden outcomes are revealed. In the simulator, replaying one realized sequence yields the *fixed-sequence effect*, while rerunning the adaptive interaction yields the *complete-loop effect*; their difference is the *additional-loop effect*.

This difference measures the additional change observed when downstream adaptation is allowed to operate relative to replaying the realized reference sequence. It is conditional on that sequence and is therefore an operational accounting of the loop, not a path-independent mediation effect.

Not all Influence is necessarily undesirable. A user may welcome a new discovery or endorse a change in taste. To make the simulator evaluation explicit, the hidden evaluator therefore uses a fixed benchmark definition of unwanted change:

$$v_t = \begin{cases} 0, & \text{if the change is user-endorsed,} \\ \max(0, \|x_t - x_0\|_1 - \delta_x) + g(a_{1:t}, m_i), & \text{otherwise,} \end{cases} \tag{2}$$

where x_t is the simulated preference state, m_i is a protected meta-preference, δ_x is a small tolerance, and g penalizes action histories that violate that meta-preference. The fixed-sequence and complete-loop effects are paired differences in this benchmark score between influence-enabled and matched no-influence worlds. We use this score as hidden simulator reference truth; it is not a general measure of human welfare, consent, or harm.

We also separate Influence from service quality. Service is measured in the complete no-influence world, so a system does not receive credit simply because it changes the user who later provides feedback. The observable audit does not access latent preference states, v_t, or either hidden effect directly. Instead, it uses ordinary feedback and the controlled preference check introduced next, and is evaluated against the hidden reference effects only after the observable comparison has been fixed.

4 Learning from Observable Interactions

The effects defined in Section 3 rely on hidden simulator state and paired counterfactual worlds, and therefore cannot be observed directly from ordinary interactions. The key question is whether an observable signal can distinguish *Influence* from *Adaptation* without access to this hidden reference truth.

Ordinary feedback is poorly suited to this task because it follows the personalized action. A higher rating may reflect better service, stronger alignment with an existing preference, or a preference that has already been changed by previous actions. We therefore introduce a *controlled preference check*: at selected interaction steps, the evaluator queries a preference dimension k_t determined by a

schedule fixed independently of the current personalized action. The resulting controlled response is an observable, noisy probe of that preference dimension rather than feedback on the item selected by the agent.

At each scheduled query, the controlled check records

$$A_t = e_t(k_t) \max(0, |r_t(k_t) - r_0(k_t)| - 0.01)\,, \tag{3}$$

where $r_t(k_t)$ is the controlled response and $e_t(k_t)$ is the observed exposure share of the queried dimension. The response term directly probes displacement along the scheduled preference coordinate, while the exposure term weights that displacement by how strongly the queried dimension has appeared in the interaction history. Only the queried dimension is assigned independently of the current personalized action, so the full score remains partly action-dependent. We therefore interpret the controlled check as a controlled preference-state probe rather than a causal estimator of Δ_{loop}. In our simulator, the controlled response is a low-noise proxy for the corresponding preference coordinate, making this a deliberately favorable state-probe setting rather than a model of arbitrary real-world self-report.

For matched pair j, we compare before–after controlled responses between the influence-enabled user and its matched no-influence user:

$$\widehat{\Delta}_j^{\text{anchor}} = (A_{j,1}^{\text{post}} - A_{j,1}^{\text{pre}}) - (A_{j,0}^{\text{post}} - A_{j,0}^{\text{pre}}). \tag{4}$$

Ordinary feedback uses the same matched before–after comparison. The goal of the controlled check is not to estimate the numerical magnitude of Δ_{loop}, but to recover its ordering across conditions from observable interactions.

To preserve the separation between observable estimation and hidden evaluation, the evaluator uses only interaction-level information, service estimates, matched-pair assignments, and controlled-check responses. It does not observe latent preference state, response-mechanism labels, v_t, or paired-world Influence scores. Observable estimates are saved before the hidden effects are revealed. This save-before-reveal boundary ensures that the observable signal is evaluated against a target it cannot directly read.

5 Experiments

5.1 Experimental Setup

Domain and simulated users. We use MovieLens-1M ratings and movie genres to construct heterogeneous simulated users and recommendation choice sets (Harper & Konstan, 2015). MovieLens provides the initial user–item data but does not contain causal labels of preference influence; preference dynamics, service utility, and benchmark-defined preference change are controlled components of our simulator. Our results therefore concern MovieLens-grounded simulated users rather than real MovieLens users.

Personalization systems. We compare two systems that differ only in whether their persistent user profile is updated. The *fixed-profile* system retains its initial structured profile throughout the interaction, whereas the *updating-profile* system revises the profile every ten interactions using recent interaction evidence. The profile contains bounded preference scores over movie genres. In the frozen held-out evaluation, profile updates are generated by Qwen3.5-9B as validated JSON operations; the language model itself is never fine-tuned or otherwise updated.

User-response mechanisms. We evaluate three simulated mechanisms through which agent actions can influence user preferences. *Assimilation* moves preferences toward recent exposure, making repeatedly exposed genres more attractive. *Satiation* instead decreases preference under repeated exposure. The *delayed-response* mechanism maintains an exposure-memory state, produces state-dependent preference change, and gradually recovers toward the initial state. These mechanisms create qualitatively different relationships between exposure, preference change, and observed feedback.

Experimental design. The frozen held-out study contains 40 simulated users arranged into 20 matched pairs, three response mechanisms, two personalization systems, five influence levels $\{0, 0.025, 0.05, 0.10, 0.20\}$, and three paired environment seeds. Each interaction trajectory contains

Table 1: **Controlled preference check versus ordinary feedback.** Panel A evaluates cross-dose ordering of hidden complete-loop benchmark-defined change across six response-mechanism–system strata. Nominal influence level is shown as a diagnostic reference because it exposes the favorable monotonic structure of the aggregate task. Panel B evaluates the harder same-dose comparison between the fixed- and updating-profile systems. Same-dose intervals resample the 20 matched pairs.

Panel A: cross-dose ordering

Signal	Macro ρ	Worst ρ	Gain vs. feedback	Worst gain
Ordinary feedback	.1333	-1.0000	.0000	.0000
Controlled check	1.0000	1.0000	.8667	.1000
Nominal influence level	1.0000	1.0000	.8667	.1000

Panel B: same-dose system ordering

Rule	Accuracy (95% CI)	Balanced accuracy (95% CI)	Mean regret (95% CI)
Controlled check	.8125 [.7417, .8792]	.6601 [.5620, .7651]	.00318 [.00163, .00515]
Ordinary feedback	.4917 [.3708, .6167]	.3622 [.2716, .4572]	.02342 [.01560, .03145]
Always fixed	.8083 [.7458, .8667]	.5000 [.5000, .5000]	.00444 [.00226, .00764]

100 steps, with the first ten used as warmup. All observable estimates are saved before hidden effects are revealed, following the information-separation protocol in Section 4.

Our experiments answer two questions. First, *can observable interactions distinguish Influence from Adaptation?* We compare the controlled preference check and ordinary feedback against hidden complete-loop preference change. The primary evaluation orders five influence levels within each of the six response-mechanism–system strata; a harder same-dose diagnostic then compares the fixed- and updating-profile systems without relying on differences in configured influence level. Second, *does self-updating recursively amplify Influence?* We compare the fixed-sequence and complete-loop effects in the prespecified delayed-response, updating-system condition at influence level 0.10, evaluating Δ_{add}against the frozen practical threshold of 0.01.

Supporting analyses. We use profile-update fallback sensitivity, a reduced Qwen3-4B run, development fixed-dose analyses, service-floor sensitivity, and save-before-reveal checks to characterize robustness and scope. These analyses do not alter the frozen held-out claims. Full simulator definitions, update rules, statistical thresholds, and artifact provenance are provided in the appendix.

5.2 Experimental Results

5.2.1 Controlled preference check vs. ordinary feedback

The frozen held-out evaluation contains all 1,020 expected fragments across 40 users in 20 matched pairs, three seeds, three response mechanisms, two systems, and five influence levels. We first ask whether an observable signal can recover the ordering of hidden complete-loop benchmark-defined change. Table 1 combines the primary cross-dose evaluation with the harder same-dose diagnostic, and Figure 3 shows the condition-level ordering result.

Cross-dose ordering. In Table 1 Panel A, across the six response-mechanism–system strata, the controlled preference check achieves a Spearman correlation of 1.0 with hidden complete-loop preference change in every stratum, compared with a macro-average correlation of 0.133 for ordinary feedback. Figure 3 shows that this difference is consistent across response mechanisms and both personalization systems. In particular, ordinary feedback reverses the hidden ordering in both satiation strata ($\rho = -1.0$), whereas the controlled check remains perfectly monotonic. Under satiation, repeated exposure reduces preference for the exposed genre, so observed feedback can move in the opposite direction from underlying preference change. This illustrates why feedback generated by the personalized action can be an unreliable signal of Influence.

This aggregate result is intentionally not treated as sufficient construct evidence. Nominal influence level also obtains perfect macro rank, showing that configured dose is itself strongly monotonic with hidden change at this level. The cross-dose experiment therefore establishes that the controlled check preserves the benchmark ordering and that ordinary feedback need not do so; it does not by

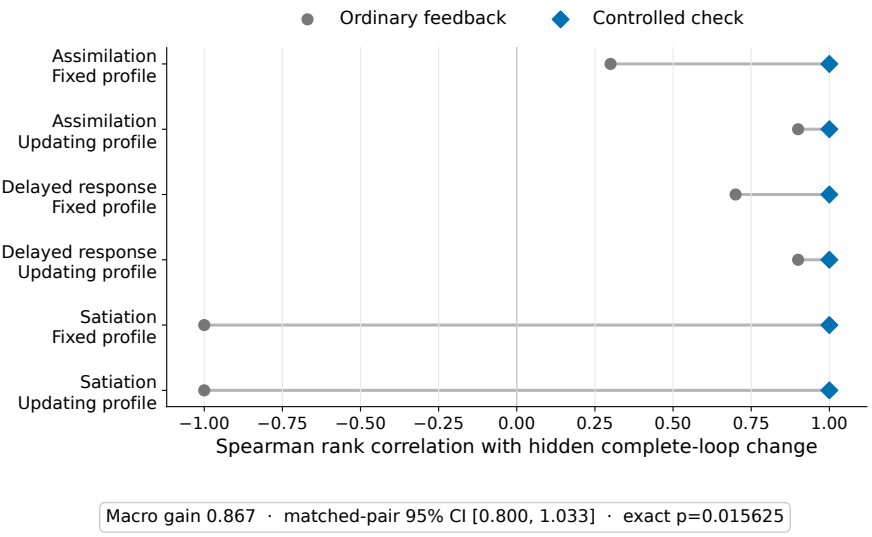


Figure 3: **The controlled preference check preserves the ordering of hidden complete-loop Influence across all six strata.** Bars show Spearman correlations across the five influence levels for each response-mechanism–system stratum. Ordinary feedback reverses the ordering under satiation, whereas the controlled check remains monotonic with hidden complete-loop change.

itself show that the check recovers variation beyond nominal dose. We therefore turn to the harder same-dose diagnostic.

Harder same-dose diagnostic. We therefore compare the fixed- and updating-profile systems under the same configured influence level. In Table 1 Panel B, among 240 non-tied matched-pair comparisons, the controlled preference check agrees with the hidden system ordering in 195 cases, or 81.25%, with a matched-pair bootstrap interval of $[73.75\%, 87.92\%]$. Raw accuracy is only slightly higher than the 80.83% obtained by always choosing the fixed-profile system. The difference is clearer in balanced accuracy: the controlled check reaches 66.0%, compared with 50.0% for the always-fixed baseline, because it identifies 19 of the 46 cases in which the updating-profile system causes less hidden change. The balanced-accuracy improvement has an interval of $[6.2, 26.5]$ percentage points, whereas the intervals for raw-accuracy and regret differences cross zero. We therefore interpret the same-dose result as evidence that the controlled check contains information about minority cases, not as broad superiority for pairwise system selection. Appendix Table 9 and Figure 8 report the full mechanism–dose breakdown.

The contrast with ordinary feedback is especially clear under satiation. Repeated exposure lowers the simulated user's preference for the exposed genre, so feedback on the selected item can move in the opposite direction from hidden preference change. Figure 4 shows this reversal directly.

Taken together, the cross-dose and same-dose results support the controlled check as an observable *ordering* signal in this simulator, not as a calibrated estimator of Influence or harm. A fixed-dose development analysis that varies hidden susceptibility and recovery gives the same qualitative conclusion beyond nominal dose (Appendix H), while the retrospective component analysis shows that controlled response displacement, rather than exposure alone, carries most of the useful signal. Table 8 summarizes these construct boundaries.

5.2.2 Does self-updating amplify Influence?

We next ask whether preference change propagates further through repeated feedback, profile updates, and subsequent actions. In the prespecified delayed-response, updating-system condition at influence level 0.10, the influenced and no-influence reruns produce different action paths for every user. Nevertheless, the complete-loop effect only slightly exceeds the fixed-sequence effect, as summarized in Table 2.

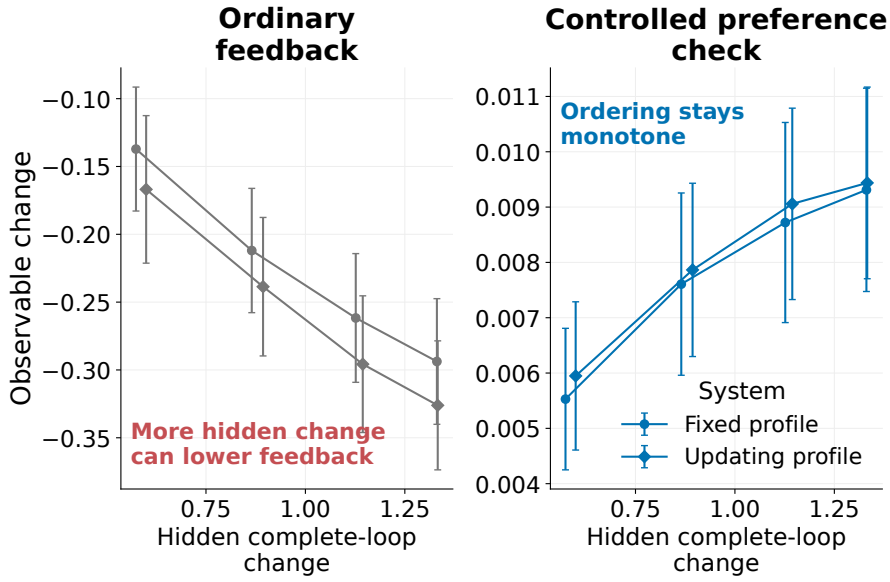


Figure 4: **Ordinary feedback can reverse the ordering of Influence under satiation.** The controlled preference check continues to follow hidden complete-loop preference change across influence levels.

Table 2: **Prespecified additional-loop test.** The influenced and no-influence trajectories diverge in action space, and the additional-loop effect is positive, but its magnitude remains below the frozen 0.01 threshold for practically meaningful amplification.

Quantity	Value
Fixed-sequence effect Δ_{fixed}	.536187
Complete-loop effect Δ_{loop}	.539141
Additional-loop effect Δ_{add}	.002954
95% CI for Δ_{add}	[.000748, .005357]
One-sided lower bound	.001084
Users with divergent action paths	100%
Practical threshold	.010

The one-sided lower bound is positive, so repeated adaptation contributes some additional benchmark-defined preference change in this condition. Its magnitude, however, is less than one third of the frozen 0.01 threshold. The data therefore do not support the stronger claim that the self-updating loop produces practically meaningful recursive amplification.

The broader diagnostic curves show why the decomposition matters. Assimilation and delayed response yield small positive additional effects in parts of the grid, whereas satiation can produce negative additional effects (Appendix Figure 9). Thus, feedback-driven profile updating can amplify or attenuate preference change depending on the simulated response dynamics. The supported conclusion is consequently narrower but more informative: changed actions, or even a large complete-loop effect, should not be interpreted as evidence of recursive amplification without comparing against the fixed-sequence counterfactual.

5.2.3 Robustness and scope

Profile-update execution. The Qwen3.5-9B updater makes 36,720 proposal decisions. Final-candidate validity is 1.0, model-only validity after one repair is 0.960975, and validator-backed exact-keep fallback occurs in 1,433 decisions (3.90%). Restricting the analysis to fragments with no fallback preserves a controlled-check macro rank of 1.0 and a gain of 0.8333 over ordinary feedback. Fallback rate is only weakly associated with updating-system action divergence ($\rho = 0.0762$) and the additional-loop effect ($\rho = -0.1074$). Full fallback diagnostics are reported in Appendix Table 11.

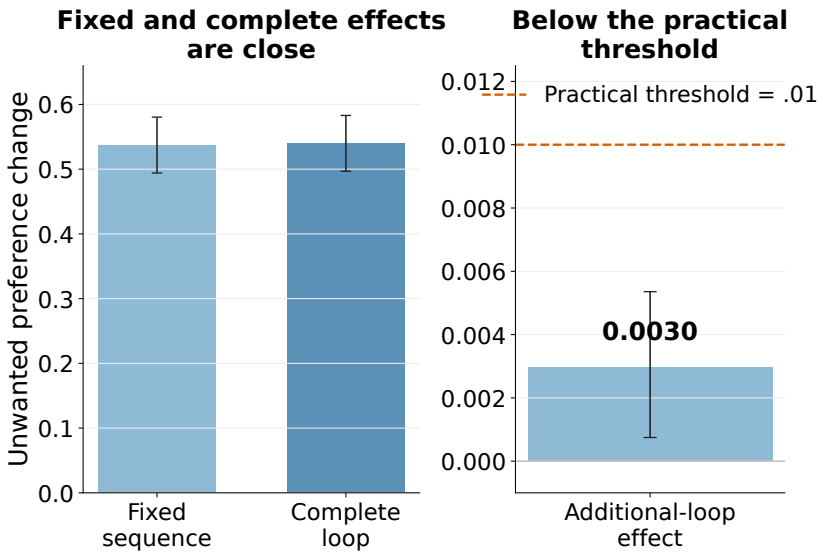


Figure 5: **Action-path divergence does not imply substantial recursive amplification.** The additional-loop effect is 0.002954, below the prespecified practical threshold of 0.01.

Updater consistency and development checks. A smaller exploratory run using Qwen3-4B shows the same aggregate ordering pattern: controlled-check macro rank is 1.0000, compared with 0.1667 for ordinary feedback, for a gain of 0.8333. Because its four matched pairs are a subset of the primary cohort and the run was scheduled after observing the main result, we treat it as consistency evidence rather than an independent replication (Appendix Table 12). Earlier fixed-dose development analyses likewise vary hidden susceptibility and recovery while holding configured influence fixed, yielding a controlled-check mean rank of 0.9903 with worst-stratum rank 0.9500, compared with 0.2431 for ordinary feedback. These analyses support the interpretation of the controlled preference check but do not alter the frozen held-out claims.

System choice is a boundary, not a contribution. At the frozen 95% service floor, the controlled check and hidden oracle both select the fixed-profile system in all 15 non-control cells, while ordinary feedback selects the updating-profile system in two high-dose cells. However, always selecting the fixed-profile system also matches the oracle in all 15 cells, and the conclusion is not robust to the service threshold: across 1,030 exact floors and 16 eligibility regimes, the simultaneous lower confidence band crosses zero and leave-one-pair-out stability fails. We therefore treat system choice only as a boundary analysis, not as evidence that the controlled check generally selects a safer personalization system. Appendix Figure 10 and Section I discuss this limitation in detail.

Observable and hidden computations remain separated. Changing the hidden violation calculation does not alter the saved controlled-check comparison, while changing controlled-response noise does not alter the hidden fixed-sequence or complete-loop effects. The save-before-reveal protocol therefore preserves the intended distinction between observable auditing and hidden simulator evaluation. Appendix E and Appendix H provide the field-level and execution checks.

6 Conclusion

Feedback produced inside a personalization loop can mix better service with a changed simulated user. In our controlled benchmark, a scheduled preference-state probe orders hidden complete-loop change substantially better than ordinary feedback across influence levels and retains useful, though imperfect, information in a harder same-dose comparison. Separately, the fixed-sequence versus complete-loop decomposition shows that fully divergent action trajectories need not imply practically meaningful recursive amplification: the additional-loop effect is positive but remains below the prespecified threshold in the delayed-response setting. These results support controlled auditing of Adaptation versus Influence while also delineating its limits; the benchmark is synthetic, the query model is favorable, and the fixed-profile baseline is structurally advantaged.

Ethics Statement

Preference influence is a normative and domain-dependent topic. We use simulated users so that paired causal comparisons are possible without experimenting on real people. The benchmark treats user-endorsed discovery differently from protected-preference violations, but this is not a universal welfare definition. Any real deployment of controlled preference checks would require informed study design, limits on user burden, privacy protection, and independent review.

Reproducibility Statement

Every component needed to rerun the frozen held-out evaluation is in the appendix: the three user-response mechanisms, the hidden benchmark-defined change score of Equation 2, and the negative controls (Appendix D); the two systems, update cadence, deterministic validator, exact-keep fallback, MovieLens-1M manifest, and configuration grid (Appendix B, Table 3); the controlled-check score and matched-pair comparison (Appendix C); the information boundary between system, observable evaluator, and hidden scorer (Table 4); and every frozen threshold beside its observed value (Table 5, Appendix F). Profile updates come from Qwen3.5-9B at temperature 0.7, top-$p = 0.8$, top-$k = 20$, and base seed 20260923, with no weight updates; Appendix H.7 reports a reduced Qwen3-4B run. The frozen run comprises 1,020 verified pair–cell–seed fragments, and observable estimates are hashed and sealed before hidden outcomes are revealed, with all digests recorded in Appendix J. Post-result diagnostics are reported separately from the frozen decisions they cannot revise (Appendix H), and the open stress tests are listed in Appendix K.

AI Use Statement

Generative AI tools assisted with literature organization, code review, and manuscript polishing. All reported result values were cross-checked against saved experiment artifacts; configuration and execution claims were checked against their corresponding manifests and systems.

References

Hal Ashton and Matija Franklin. Solutions to preference manipulation in recommender systems require knowledge of meta-preferences. *arXiv preprint arXiv:2209.11801*, 2022.

Muhammad Rafay Azhar, Yuhang Zhou, Gilbert Jiang, Yuchen Wang, Rahul Sharma, Matthew DeSousa, Jiayi Liu, Xin Guo, Lizhu Zhang, and Xiangjun Fan. Coral: An llm-native harness for production recommender systems. *arXiv preprint arXiv:2609.02730*, 2026.

Micah D Carroll, Anca Dragan, Stuart Russell, and Dylan Hadfield-Menell. Estimating and penalizing induced preference shifts in recommender systems. In *International Conference on Machine Learning*, pp. 2686–2708. PMLR, 2022.

F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19, 2015.

Atoosa Kasirzadeh and Charles Evans. User tampering in reinforcement learning recommender systems. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 58–69, 2023.

David Krueger, Tegan Maharaj, and Jan Leike. Hidden incentives for auto-induced distributional shift. *arXiv preprint arXiv:2009.09153*, 2020.

Ming Li, Dai Li, Xuying Ning, Bo Sun, Rui Li, Yi Zhang, Silvia Gong, Xuan Cao, Cornelia Carapcea, Qunshu Zhang, et al. Auto-recsys: Harnessing autonomous research agents for industry-scale recommender system. *arXiv preprint arXiv:2609.10922*, 2026.

Kaiqu Liang, Julia Kruk, Shengyi Qian, Xianjun Yang, Shengjie Bi, Yuanshun Yao, Shaoliang Nie, Mingyang Zhang, Lijuan Liu, Jaime Fernández Fisac, et al. Learning personalized agents from human feedback. *arXiv preprint arXiv:2602.16173*, 2026.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *International conference on machine learning*, pp. 7599–7609. PMLR, 2020.

Cheng Qian, Zuxin Liu, Akshara Prabhakar, Zhiwei Liu, Jianguo Zhang, Haolin Chen, Heng Ji, Weiran Yao, Shelby Heinecke, Silvio Savarese, et al. Userbench: An interactive gym environment for user-centric agents. *arXiv preprint arXiv:2507.22034*, 2025.

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.

Liyan Tang, Cyrus Rashtchian, Chun-Sung Ferng, Andrew Tomkins, Da-Cheng Juan, and Tu Vu. Wikiskill: Compiling agent experience into persistent knowledge for skill evolution. *arXiv preprint arXiv:2608.27454*, 2026.

Zhen Tao, Riwei Lai, Chenyun Yu, Weixin Chen, Li Chen, Beibei Kong, Lei Cheng, Chengxiang Zhuo, Zang Li, and Qingqiang Sun. Sager: Self-evolving user policy skills for recommendation agent. *arXiv preprint arXiv:2604.14972*, 2026.

Haochen Wang, Yi Wu, Daryl Chang, Li Wei, and Lukasz Heldt. Self-evolving recommendation system: End-to-end autonomous model optimization with llm agents. *arXiv preprint arXiv:2602.10226*, 2026.

Weizhi Zhang, Xinyang Zhang, Chenwei Zhang, Liangwei Yang, Jingbo Shang, Zhepei Wei, Henry Peng Zou, Zijie Huang, Zhengyang Wang, Yifan Gao, et al. Personaagent: Bridging memory and action for personalized llm agents. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 26421–26439, 2026.

A LIMITATIONS

Our study is conducted in a controlled simulation and should therefore be interpreted as an auditing benchmark rather than a calibrated measure of real-user harm, welfare, or consent. The controlled preference check uses a deliberately favorable low-noise proxy for simulated preference state, while the fixed-profile baseline begins from the true initial preference state and is therefore structurally advantaged. MovieLens provides users and items, but the preference dynamics, endorsement rules, and hidden benchmark-defined change are simulated. Extending these findings to real users would require substantially noisier observations, genuine preference drift, competitive adaptive systems, and separately designed human studies.

B COMPLETE EXPERIMENTAL SETUP

B.1 MOVIELENS MANIFEST

The supplied MovieLens-1M manifest records 1,000,209 ratings, 6,040 users, 3,883 movies, 2,945 eligible users, a development/validation/test split, one example initial-preference vector, and the archive SHA-256 reported in Section J. It does not document further preprocessing or provide causal influence labels.

B.2 PERSISTENT PROFILE AND UPDATE FORMAT

The fixed system keeps its initial profile. The updating system proposes a profile change after steps 10, 20, ..., 90. From the last ten interactions, the proposal and its observable replay-gain check both use the older six; the newest four are excluded from that update decision. Qwen3.5-9B receives the current 18 genre scores, shown genres, ordinary feedback, the declared meta-preference, and an exact evidence-support map. It may change one to three genres by at most 0.05 in an evidence-supported direction.

Generation uses at most 8,192 prompt tokens and 384 new tokens, with temperature 0.7, top-$p = 0.8$, top-$k = 20$, and base seed 20260923. A deterministic validator checks the JSON schema, parent-profile hash, bounds, directions, exact safe values, and cited interaction IDs. An invalid sampled proposal receives one greedy repair; a second failure yields a validator-backed exact `keep`. The candidate is applied only when its replay over those same six interactions improves the frozen update gate. Model weights never change.

C OPERATIONAL DEFINITIONS AND OBSERVABLE COMPARISON

At interaction t, the system chooses action a_t from persistent profile h_t and choice set c_t. The simulated user state x_t, feedback y_t, and profile update follow

$$a_t \sim \pi(\cdot \mid h_t, c_t), \qquad x_{t+1} = F(x_t, a_t, \epsilon_t), \tag{5}$$

$$y_t = Y(x_t, a_t, \eta_t), \qquad h_{t+1} = U(h_t, \tau_t) \quad \text{at update times.} \tag{6}$$

The updater changes the stored profile, not model weights. All averages use the post-warmup window $\mathcal{T}_{\text{post}} = \{11, \ldots, 100\}$.

The fixed-sequence and complete-loop effects are

$$\Delta_{\text{fixed}}(s) = \frac{1}{|\mathcal{T}_{\text{post}}|} \sum_{t \in \mathcal{T}_{\text{post}}} \left[v_t^{\text{influence|fixed}} - v_t^{\text{no influence|fixed}} \right], \tag{7}$$

$$\Delta_{\text{loop}}(s) = \frac{1}{|\mathcal{T}_{\text{post}}|} \sum_{t \in \mathcal{T}_{\text{post}}} \left[v_t^{\text{influence|rerun}} - v_t^{\text{no influence|rerun}} \right]. \tag{8}$$

The additional-loop effect is

$$\Delta_{\text{add}}(s) = \Delta_{\text{loop}}(s) - \Delta_{\text{fixed}}(s). \tag{9}$$

The influence-enabled complete loop supplies the fixed reference sequence, so this decomposition is conditional on that path.

Table 3: **Development, delivery-check, and frozen held-out configurations.** Development ordering and retrospective component analyses use a scripted updater. The reduced Qwen run tests the distributed execution path and is not a scientific result; the complete Qwen held-out run supplies the primary evidence.

Component	Setting
Domain	MovieLens-1M-grounded simulation
Development updater	Scripted profile update
Held-out updater	Frozen Qwen3.5-9B
System roles	Fixed profile; updated profile
Update cadence	Every 10 interactions
Development/held-out trajectory	100 steps; 10-step warmup
Response mechanisms	Assimilation; satiation; delayed response with memory
Influence levels	0, .025, .05, .10, .20
Environment seeds	3 paired seeds
Initial development cohort	10 users; 90,000 public steps
Same-setting development cohort	40 users in 20 matched pairs
Same-setting public artifact	1,368,000 steps
Component analysis	Replay of the same 1,368,000 public rows
Development floor sensitivity	1,003 floors; 20 matched pairs
Held-out cohort	40 users in 20 matched pairs
Held-out grid	17 cells; 1,020 verified pair–cell–seed fragments
Held-out public artifact	408,000 interaction steps
Held-out diagnostics	1,030 floors; 16 eligibility regimes
Qwen delivery check	16 users in 8 matched pairs; 64 fragments
Qwen weight update	None

Service quality is measured in the complete no-influence world:

$$\widehat{Q}_c(s) = \frac{1}{|\mathcal{T}_{\text{post}}|} \sum_{t \in \mathcal{T}_{\text{post}}} u_{t,c}^{\text{no influence|rerun}}. \tag{10}$$

For equations and artifact names, the controlled preference check is called a controlled anchor. At scheduled preference dimension k_t, it records

$$A_t = e_t(k_t) \max(0, |r_t(k_t) - r_0(k_t)| - 0.01), \tag{11}$$

where $e_t(k_t)$ is observed exposure share. For environment seed s, the scheduled genre is $k_t = (7s + t - 1) \bmod 18$. The baseline and current responses are $r_0(k) = \gamma_{ik} x_{0,k}$ and $r_t(k) = \gamma_{ik} x_{t,k} + 0.02\eta_t$, where $\gamma_{ik} = 0.80 + 0.05((i + 3k) \bmod 9)$. Only k_t is assigned independently of the current action. Because $e_t(k_t)$ includes exposure after that action, A_t is action-dependent. Before clipping, the additive query-noise standard deviation is 0.0003.

For matched pair j, the observable comparison is

$$\widehat{\Delta}_j^{\text{anchor}} = (A_{j,1}^{\text{post}} - A_{j,1}^{\text{pre}}) - (A_{j,0}^{\text{post}} - A_{j,0}^{\text{pre}}). \tag{12}$$

Ordinary feedback uses the same matched before–after form. The anchor is evaluated as an ordering signal; it is not a calibrated estimate of Δ_{loop}.

D User Simulation and Hidden Evaluation

D.1 Three response mechanisms

Immediate assimilation. Recent exposure moves the simulated preference vector toward the exposed genre mixture. This mechanism can make repeated exposure increasingly attractive.

Satiation. Repeated exposure lowers the short-term preference for the exposed genre. Observed feedback can therefore move in the opposite direction from the assimilation setting even at the same nominal influence level.

Table 4: **Who can see which fields.** A check mark means the component may read the field. The observable estimate and chosen system are saved before the final column is opened.

Information	Personalization system	Observable evaluator	Hidden scorer
Interaction history and offered items	✓	✓	✓
Personalized actions and ordinary feedback	✓	✓	✓
Current profile contents	✓		✓
Profile version and update timing	✓	✓	✓
Matched-pair assignment and pre-period outcomes		✓	✓
Service utility		✓	✓
Controlled-anchor responses		✓	✓
Unwanted-change score and components			✓
Latent user state and transition rule			✓
Exact paired-world outcomes			✓
Fixed-sequence and complete-loop effects			✓
Final hidden ranking and regret			✓

Delayed response with memory. An exposure-memory state accumulates over time, changes response strength, and recovers toward the initial state when exposure stops. The mechanism is delayed and state-dependent, so it is structurally different from an immediate linear response.

D.2 Unwanted change and endorsement

The hidden evaluator computes a state-and-action violation:

$$v_t = \begin{cases} 0, & \text{if the change is user-endorsed,} \\ \max(0, \|x_t - x_0\|_1 - \delta_x) + g(a_{1:t}, m_i), & \text{otherwise.} \end{cases} \tag{13}$$

Here m_i is a protected meta-preference, δ_x is a small tolerance, and g penalizes action histories that violate that meta-preference. Neither v_t nor its components are available to the personalization system or the observable evaluator. The controlled anchor is computed from a different response channel and is never inserted into v_t.

D.3 Negative controls

The zero-influence control disables user change caused by the chosen action. The matched outside-drift control applies recipient-independent drift with the same recipient-specific transition magnitude. The donor control uses actions from a different development user, again matched to the recipient's transition magnitude. These controls test whether state movement or action magnitude alone reproduces the observable comparison. Donor users come only from the development partition.

E Information Available to Each Component

Condition codes exposed to the observable evaluator reveal neither the response mechanism nor the influence level or hidden ordering. The source defines the public-field allowlist, and the saved estimate and seal record the output and hashes. Hidden truth is committed before sealing and revealed only after seal verification.

F Held-Out Statistical Rules

F.1 Held-out ranking rule

The six mechanism–system conditions receive equal weight. Average ranks are used for ties. A condition whose hidden complete-loop effect is constant across all five influence levels is a data-adequacy failure. If an observable signal is constant while the hidden effect varies, its Spearman correlation is defined as zero.

For each arm and seed, public pre- and post-period values average steps 1–10 and 11–100. We then average the three seed summaries and form the matched treatment–control difference. For

Table 5: **Frozen decision rules and observed values.** A failed row for the extra feedback-loop effect does not invalidate the decomposition; it prevents a positive practical-amplification claim. The system-choice row applies only at the frozen 95% service floor; the broader post-result floor-family robustness check fails separately.

Criterion	Frozen threshold	Observed value	Decision
System choice at the frozen service floor			
Controlled-check regret is lower than feedback regret	Check $<$ feedback	0.0000 vs. 0.0101	Pass
Ordering complete-loop change			
Mean controlled-check rank correlation	≥ 0.80	1.0000	Pass
Mean gain over ordinary feedback	≥ 0.20	0.8667	Pass
One-sided paired sign-flip test	$p \leq 0.05$	$p = 0.0156$	Pass
Worst condition gain	≥ -0.20	0.1000	Pass
Cluster-bootstrap lower bound for mean gain	> 0	0.8000	Pass
Largest zero-influence controlled-check gap	≤ 0.01	0.0000	Pass
Largest matched-control controlled-check gap	≤ 0.01	0.0004	Pass
Observable and hidden computations remain separated	All required checks pass	All passed	Pass
Extra feedback-loop effect			
Users with diverging action paths	$\geq 25\%$	100.0%	Pass
One-sided lower bound for the additional effect	> 0	0.0011	Pass
Mean additional effect	≥ 0.01	0.0030	Fail

each user and seed, the hidden complete-loop effect is factual minus complete no-influence mean score over steps 11–100. A dose-level cell averages that effect over 40 users and three seeds. Each condition-level Spearman correlation compares the five dose-level public and hidden cell means.

The held-out ranking claim passes only if all of the following hold:

1. mean anchor Spearman correlation $\bar{\rho}_A \geq 0.80$;
2. mean gain over feedback $G \geq 0.20$;
3. exact one-sided paired sign-flip $p \leq 0.05$ over the six condition-level gains;
4. worst condition gain at least -0.20;
5. the matched-pair cluster-bootstrap 95% lower bound for G is above zero;
6. absolute zero, exogenous, and donor control gaps are at most 0.01; and
7. both separation checks pass.

The sign-flip test was part of the frozen rule, but the six conditions are not independent samples: they share users, seeds, and simulator design. We therefore treat its value as a descriptive check that all six gains have the same direction, not as population-level uncertainty. Sampling uncertainty is summarized by 2,000 matched-pair bootstrap replicates with seed 20260923. Each replicate samples the 20 oriented pairs with replacement and retains both members and all of their seeds, doses, mechanisms, and systems; matching, orientation, and thresholds are not recomputed. Average ranks handle ties. Constant hidden truth is a data-adequacy failure and a constant observable signal receives correlation zero. The frozen implementation assigns sentinel values -1 for a degenerate check correlation and -2 for a degenerate gain; no observed replicate was degenerate.

F.2 Additional loop effect

A positive feedback-loop amplification claim requires all three conditions in the delayed updating setting:

1. the influence-enabled and no-influence reruns choose different actions for at least 25% of users;

Table 6: **Earlier validation matched-control contrasts.** Each estimate is below its prespecified practical threshold.

Contrast	Estimate	Lower bound	Target
Influence − zero, adverse promotion	.0199	−.0264	.10
Influence − exogenous, harmful promotion	.0130	.0000	.05
Influence − donor, harmful promotion	.0222	.0093	.05

2. the user-cluster one-sided 95% lower bound for $\Delta_{\text{add}}$is above zero; and
3. the mean $\Delta_{\text{add}}$is at least 0.01.

The 0.01 value was frozen as a one-percentage-point decision threshold on the benchmark score before revealing held-out truth. It is not calibrated to human welfare. If any condition fails, the paper reports Δ_{fixed}, Δ_{loop}, and their difference without a positive amplification claim.

F.3 System choice

This rule is the direct test of whether observable interactions help choose the system causing less benchmark-defined change while preserving service quality. For each setting c, the observable service-quality estimate defines

$$\mathcal{S}_{\text{feasible}}(c) = \left\{ s : \widehat{Q}_c(s) \geq 0.95 \max_{s'} \widehat{Q}_c(s') \right\}. \tag{14}$$

Each comparison selects a system from this same set. The anchor comparison selects the eligible system with the lower anchor gap; ordinary feedback selects the eligible system with the higher feedback gap. Hidden complete-loop regret is

$$\text{Regret}_c(s) = \Delta_{\text{loop},c}(s) - \min_{s' \in \mathcal{S}_{\text{feasible}}(c)} \Delta_{\text{loop},c}(s'). \tag{15}$$

The controlled-anchor choice improves the pairwise decision only if it has strictly lower regret than the ordinary-feedback choice. With only one updating system and one fixed control, this is a pairwise demonstration, not a general result over personalization architectures.

G Earlier Validation Results

G.1 Matched-control study

The earlier matched-control evaluation contains 2,699 valid candidates from 2,700 opportunities. All 1,620 newly generated control candidates are valid. Gain caused by the chosen actions is exactly zero in the no-influence control, and the largest recipient-specific transition-matching error is below 5×10^{-16}. These integrity results support interpretation of the negative contrasts in Table 6; they do not turn the failed thresholds into positive evidence.

G.2 Validation on a hashed candidate pool

The user-cluster bootstrap interval for bounded minus temporal replay is $[-0.0164, 0.0899]$ for adverse promotion. The interval for bounded minus feedback-only is $[-0.1950, -0.0919]$, showing lower adverse promotion than feedback-only, but the beneficial-promotion difference is $[-0.1010, -0.0333]$. The method therefore trades away beneficial retention and misses the study's joint criterion.

H Development Ranking and Execution Evidence

Milestone labels are retained only for artifact lookup: M11 is the frozen ranking and delivery series, M13 the same-setting test, M14 the component replay, M15 the service-floor sensitivity, M17 the first held-out diagnostics, and M18 the reviewer-requested baselines and leave-one-out checks.

Table 7: **Validation results on one hashed candidate pool.** Lower adverse-candidate promotion and higher beneficial-candidate promotion are preferred. The bounded rule does not reduce adverse promotion relative to temporal replay and loses more than the allowed five points of beneficial promotion relative to feedback-only. The final two rows use the raw artifact identifiers because the supplied evidence does not establish a stronger semantic interpretation.

Rule	Adverse↓	Beneficial↑	Promote	Service gain	Coverage	Probes
Feedback-only	.8565	.9368	.8576	.00340	1.000	0
Temporal replay	**.6736**	.8421	.6698	**.00470**	1.000	0
Bounded rule	.7130	.8684	.7370	.00390	1.000	8
Hard-audit diagnostic (`hard_audit`)	.3658	.3789	.3715	.00395	.787	8
Oracle (`oracle`)	.0000	1.0000	.1763	.02794	1.000	0

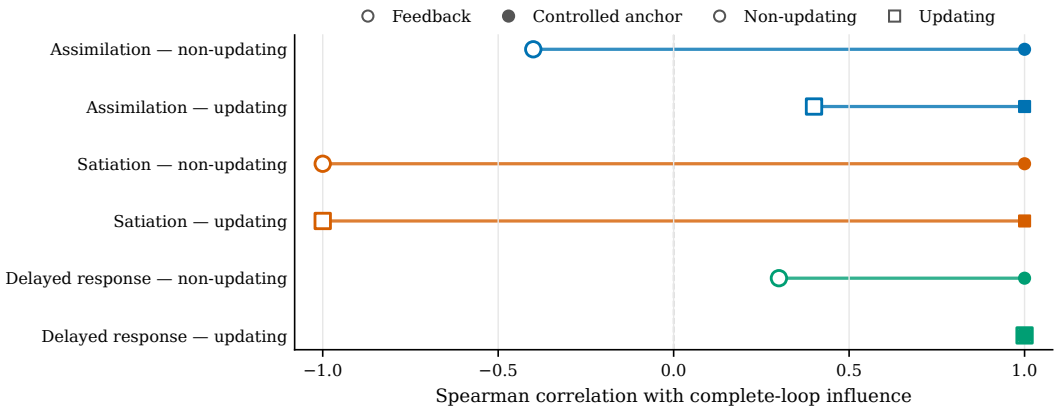


Figure 6: **Development-only comparison across all six mechanism–system conditions.** Each line joins the Spearman correlation for ordinary feedback and for the controlled anchor against the hidden complete-loop effect across five influence levels. Circles denote the non-updating system and squares the scripted updating system; color denotes the simulated user-response mechanism. The anchor reaches 1.0 in every condition, while feedback ranges from -1.0 to 1.0. This descriptive development figure does not evaluate the held-out success rule.

H.1 All six development conditions

The condition-level ordinary-feedback and controlled-anchor rank associations are:

$$\begin{aligned} \text{Assimilation, fixed:} &\quad -0.4 \to 1.0, \\ \text{Assimilation, updating:} &\quad 0.4 \to 1.0, \\ \text{Satiation, fixed:} &\quad -1.0 \to 1.0, \\ \text{Satiation, updating:} &\quad -1.0 \to 1.0, \\ \text{Delayed hysteresis, fixed:} &\quad 0.3 \to 1.0, \\ \text{Delayed hysteresis, updating:} &\quad 1.0 \to 1.0. \end{aligned}$$

These are the values visualized in Figure 6. The mean gain is 1.1167, the exact one-sided sign-flip value is 0.03125, the matched-pair bootstrap interval is $[0.9333, 1.3000]$, and the worst condition gain is zero.

H.2 Same-setting development test

The additional same-setting run (called fixed-dose in the saved artifacts) contains 40 users forming 20 matched pairs, 1,368,000 public steps, and 12 mechanism–system–dose rank conditions. Within each condition, the comparison spans nine hidden combinations of user susceptibility and recovery. The controlled-check mean rank correlation is 0.9903, its worst condition is 0.9500, and its matched-pair 95% interval is $[0.9875, 0.9958]$. The equal-condition feedback mean is 0.2431.

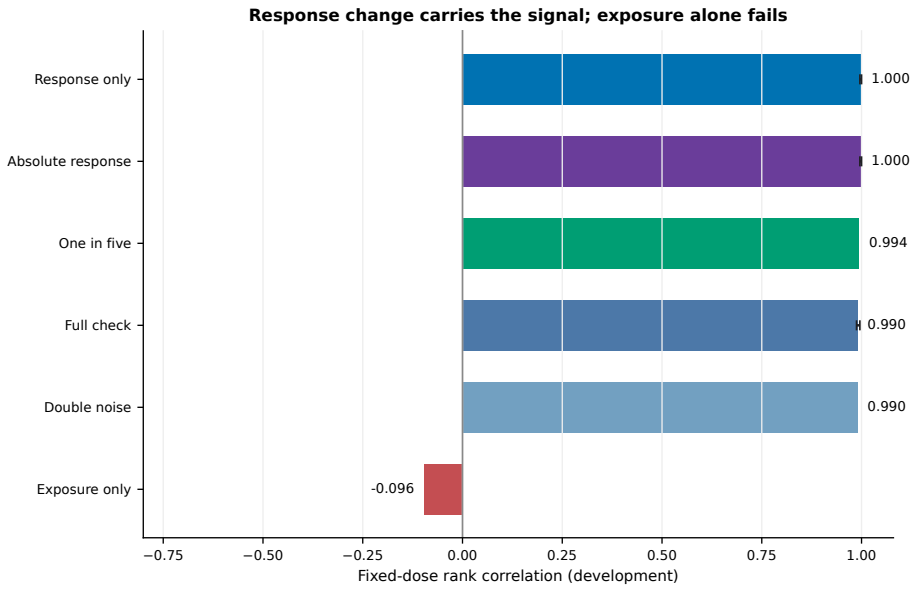


Figure 7: **The controlled response change carries the development signal.** Response-only and lower-frequency variants retain strong fixed-dose ordering, while exposure alone fails. M14 is a retrospective development analysis and cannot revise the frozen held-out result.

All eight predefined M13 checks pass. In particular, 720 fixed-system groups checked across nine hidden-factor variants have identical actions and exposure, the zero-control gap is zero, and the estimate is saved and hashed before hidden outcomes are revealed. Under the common service-quality rule, mean complete-loop regret is 0.00358 for the controlled-check choice and 0.03853 for the feedback choice.

The target-component diagnostic is selective rather than uniformly positive: the controlled check ranks simulated preference-state movement with mean correlation 0.9903, but ranks the action-history concentration component with mean correlation 0.1861. This supports the narrower preference-movement claim used in the main text.

H.3 Retrospective component analysis

M14 was designed after the M13 result and is labeled as an additional development component comparison. It replays all 1,368,000 public rows from M13. Action, profile-version, update-count, feedback, utility, and frozen anchor comparisons have zero mismatches, and the saved summary-reproduction check passes.

Response change without exposure weighting or the dead zone has mean and worst rank correlation 1.0, while exposure alone has mean -0.0958 and worst -0.7167. The paired mean difference "full minus response-only" is -0.00773 with 95% interval $[-0.0125, -0.00139]$. For the updating system, the paired mean difference "full minus exposure-only" is 1.1100 with interval $[0.6972, 1.5167]$.

Assessing only every fifth scheduled check retains mean rank 0.9944 and worst rank 0.9667. Doubling controlled-response noise leaves the frozen full-check result at mean 0.9903 and worst 0.9500. These comparisons explain the development artifact but cannot alter or rescue the frozen M11 or held-out endpoints.

Table 8: **What the aggregate ordering does and does not establish.** Cross-dose ordering is favorable because nominal influence level is strongly monotonic with hidden change. Fixed-dose development and same-dose held-out diagnostics probe whether the controlled check carries information beyond that shortcut.

Question	Comparison	Result	Interpretation
Can dose alone order the aggregate benchmark?	Five influence levels within each held-out mechanism–system stratum	Controlled check 1.000; nominal dose 1.000	Aggregate success alone does not rule out a dose shortcut.
Does the check track hidden variation when dose is fixed?	12 development conditions with nine hidden factor settings	Check .9903 [.9875, .9958]; feedback .2431	Strong development evidence beyond nominal dose.
Can it order two systems at the same dose?	240 non-tied held-out comparisons	195/240 = 81.25%, pair-bootstrap CI $[0.7375, 0.8792]$	Useful but imperfect construct evidence; not calibrated causal estimation.

H.4 Retrospective service-floor sensitivity

M15 was designed after M13 and is labeled *post-result retrospective development sensitivity*. It evaluates 1,003 service-quality floors, including the regular grid from zero to one in increments of 0.001, the exact eligibility breakpoint, and its midpoint. The analysis reproduces the M13 result at the frozen 0.95 floor to numerical tolerance.

At all 971 floors where the controlled-check and feedback rules can make different choices, the controlled rule has lower descriptive hidden regret; the advantage ranges from 0.0120 to 0.0350. It never selects lower estimated service quality. Near the single effective eligibility breakpoint at about 0.9682, however, omitting one of the 20 matched pairs can remove the comparison. An uncertainty range designed to cover the complete floor curve includes zero throughout. We therefore classify this sensitivity result as mixed rather than confirmatory.

H.5 Held-out construct, fallback, and floor diagnostics

M17 is a post-result artifact-only analysis of the frozen held-out run. It does not change any primary decision. Nominal dose and hidden fixed-path truth both have macro Spearman correlation 1.0, equal to the controlled check. At the same dose, controlled-check and hidden system order agree on 195 of 240 non-tied pair-level comparisons; the matched-pair bootstrap interval is $[0.7375, 0.8792]$. The mean within-cell pair-rank interval is $[0.3364, 0.7458]$.

Leaving out one influence level at a time preserves the descriptive aggregate ordering, as shown in Table 10. This post-result stress test reports point estimates only and does not replace matched-pair inference.

The deterministic exact-keep fallback rate is 0.039025. Restricting the rank analysis to zero-fallback fragments retains controlled-check macro rank 1.0, gain 0.8333, and worst gain 0.2. Fallback rate has Spearman association 0.0762 with action divergence and -0.1074 with the additional effect. These are nonrandomized sensitivity summaries.

The saved path-versus-loop curves also show that the additional-loop effect is not uniformly positive. Assimilation and delayed response add small positive effects in parts of the grid, while satiation attenuates change (Figure 9). This retrospective pattern explains why action divergence alone cannot establish practical amplification.

The held-out floor analysis evaluates 1,030 exact floors, 16 eligibility regimes, 10,000 matched-pair bootstrap replicates, and 20 leave-one-pair-out fits. At floor 0.95, controlled regret is zero and feedback regret is 0.010111, but the simultaneous lower band is -0.010668. The advantage is not positive in every regime where choice remains, and leave-one-pair-out stability fails. The diagnostic classification is *failure*.

Table 9: **Post-result same-dose system diagnostics.** Panel A pools 240 non-tied matched-pair comparisons. Intervals resample the 20 matched pairs and retain their mechanisms, doses, and seeds. Panel B shows raw accuracy and the number of 20 pairs where the updating system has lower hidden complete-loop change.

Panel A: pooled diagnostic

Rule	Accuracy (95% CI)	Balanced accuracy (95% CI)	Mean regret (95% CI)
Controlled check	.8125 [.7417, .8792]	.6601 [.5620, .7651]	.00318 [.00163, .00515]
Ordinary feedback	.4917 [.3708, .6167]	.3622 [.2716, .4572]	.02342 [.01560, .03145]
Always fixed	.8083 [.7458, .8667]	.5000 [.5000, .5000]	.00444 [.00226, .00764]
Always updating	.1917 [.1333, .2542]	.5000 [.5000, .5000]	.04449 [.03713, .05199]

Panel B: mechanism and dose breakdown

Mechanism	Dose	Updating wins	Check acc.	Feedback acc.	Fixed acc.
Assimilation	.025	3	.90	.50	.85
Assimilation	.050	2	.90	.50	.90
Assimilation	.100	0	.90	.40	1.00
Assimilation	.200	1	1.00	.30	.95
Delayed	.025	3	.80	.50	.85
Delayed	.050	4	.90	.40	.80
Delayed	.100	3	.85	.35	.85
Delayed	.200	1	.90	.40	.95
Satiation	.025	4	.65	.70	.80
Satiation	.050	7	.70	.70	.65
Satiation	.100	7	.65	.65	.65
Satiation	.200	11	.60	.50	.45

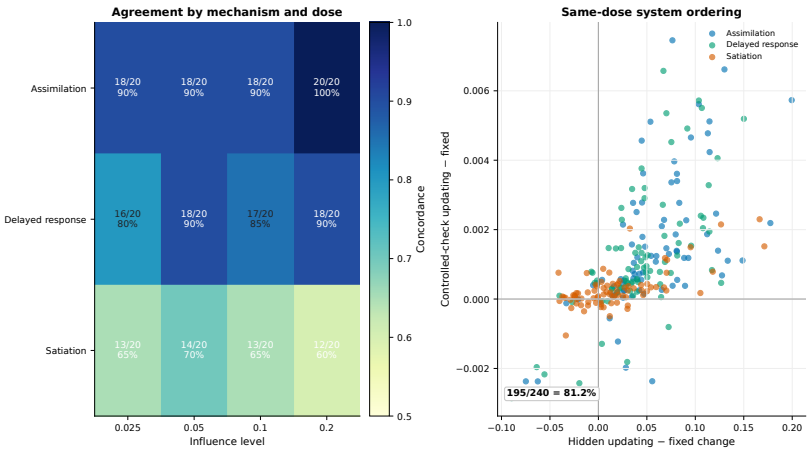


Figure 8: **Same-dose evidence is useful but imperfect.** The left panel localizes agreement by response mechanism and influence level. The right panel compares the controlled-check difference between the updating and fixed systems with the corresponding hidden complete-loop difference. The 240 comparisons reuse matched pairs and are not independent samples.

H.6 Distributed execution check

The eight-rank delivery check contains 64 atomic fragments, eight matched pairs, 16 users, 7,680 public interaction steps, and 512 Qwen update attempts. All eight ranks produce updates. Model-only validity is $495/512 = 0.9668$. Seventeen attempts use the exact-keep fallback, final-candidate validity is 1.0, and no deterministic change fallback occurs. The maximum absolute zero-control anchor gap is zero; the maximum absolute exogenous or donor control gap is 0.0007767. The recorded terminal guardrail confirms completion of the sealed observable artifact and truth-reveal sequence.

Table 10: **Descriptive leave-one-dose-out stability.** This post-result stress test reports point estimates only and does not replace the frozen matched-pair inference.

Grid	Check ρ	Feedback ρ	Gain	Worst gain
Full five-dose grid	1.0000	0.1333	0.8667	0.1000
Omit 0	1.0000	-0.0667	1.0667	0.2000
Omit 0.025	1.0000	0.0333	0.9667	0.2000
Omit 0.05	1.0000	0.1333	0.8667	0.2000
Omit 0.1	1.0000	0.2000	0.8000	0.0000
Omit 0.2	1.0000	0.3000	0.7000	0.0000

Table 11: **Fallback sensitivity.** Restricting to fragments with no deterministic fallback preserves the controlled-check ordering. Fallback strata are descriptive, not randomized.

Subset / diagnostic	Value 1	Value 2
All rank fragments	Check rank 1.0000	Gain 0.8667
Zero-fallback rank fragments	Check rank 1.0000	Gain 0.8333
Zero fallback	435 fragments	Additional 0.0005
One or more fallbacks	585 fragments	Additional -0.0037
Fallback rate correlations	Divergence 0.076	Additional -0.107

The check's reduced scientific summaries are intentionally not used as evidence; its purpose is delivery validation.

H.7 Reduced run with a smaller profile updater

We also run the same audit with a smaller Qwen3-4B profile updater on eight users in four matched pairs. This reduced run preserves all three response mechanisms, five influence levels, three seeds, and the 100-step horizon. All 204 fragments and 81,600 public interaction steps complete. Final-candidate validity is 1.0; one of 7,344 proposals uses the exact-keep fallback. The largest matched-control gap is 0.005137.

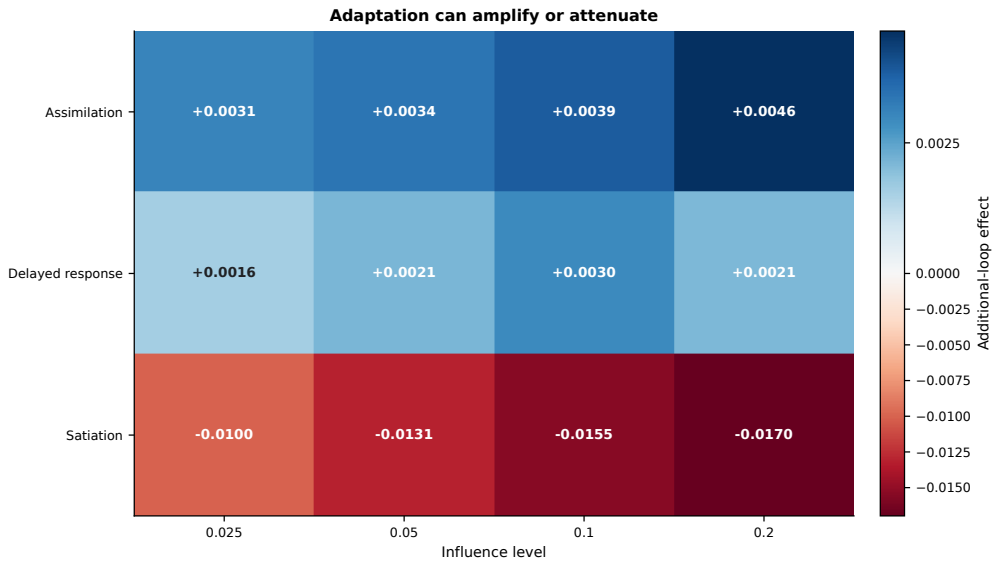


Figure 9: **Profile updating can amplify or attenuate change.** Cells show the additional-loop effect across response mechanisms, systems, and influence levels. The pattern is descriptive and cannot replace the frozen practical-magnitude gate.

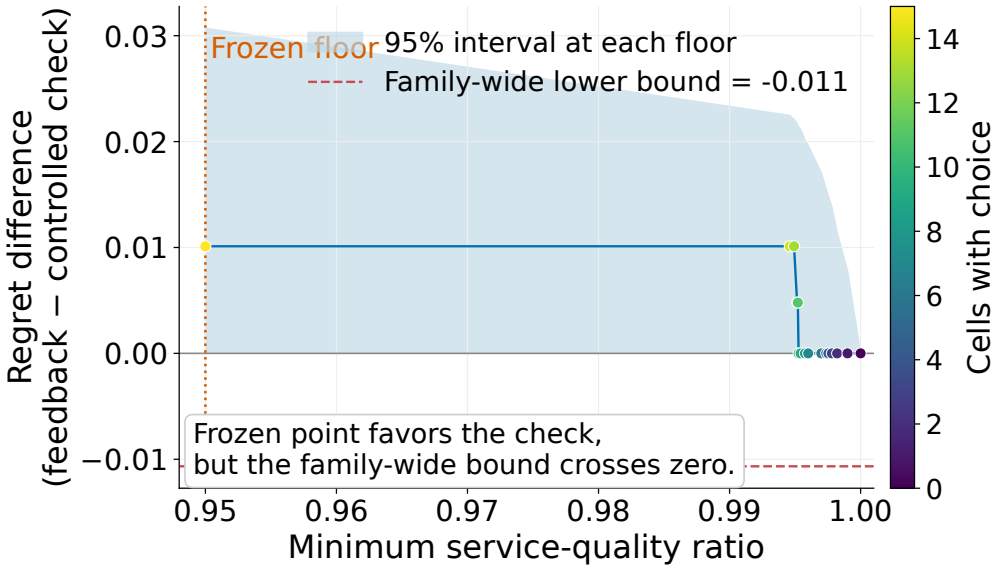


Figure 10: **System-choice conclusions depend on the service threshold.** The favorable result at the frozen 95% floor does not remain robust across the broader family of eligibility regimes. This analysis is treated as a boundary on system-selection claims rather than a positive contribution.

The smaller updater therefore repeats two qualitative findings: the controlled check orders complete-loop change more reliably than ordinary feedback, and the additional-loop effect remains below the 0.01 practical threshold. The controlled check is not oracle-perfect in this small run, but its mean selection regret is lower than feedback's. These are supporting results only; eight simulated users cannot establish broad robustness across models or user populations.

Table 12: **Exploratory consistency under a smaller updater.** The 4B result reuses four primary matched pairs and is not an independent replication.

	Qwen3.5-9B	Qwen3-4B
Users / matched pairs	40 / 20	8 / 4
Controlled-check macro rank	1.0000	1.0000
Feedback macro rank	0.1333	0.1667
Gain over feedback	0.8667	0.8333
Gain 95% CI	[0.8000, 1.0333]	[0.7333, 1.1333]
Mean additional-loop effect	0.0030	0.0062
Practical amplification claim	No	No

I Discussion

The controlled check is a favorable preference-state probe, not a calibrated harm measure. The controlled response is a deliberately low-noise proxy for a scheduled simulated preference coordinate (Section 4). In the aggregate held-out evaluation, nominal dose and hidden fixed-sequence truth also rank perfectly, so the cross-dose result alone cannot establish construct validity beyond the benchmark's monotonic dose structure. The same-dose analysis in Table 9 provides harder evidence but remains imperfect. Retrospective component analysis further shows that response displacement alone has mean and worst rank 1.0, while exposure alone is weak; this supports the narrower interpretation of the method as a clean preference-state probe rather than a general estimator of unwanted change, consent, welfare, or harm.

The fixed-profile baseline is structurally favored. The fixed profile equals the true initial simulated preference state, and the no-influence world remains at that state. The updating system therefore starts from an oracle-like profile and can perturb it. The benchmark consequently tests detection of departures from a correct stable profile, not learning under initial profile error or genuine preference drift. At the frozen service floor, always fixed matches the complete-loop oracle, and the broader floor-family analysis fails its robustness gates (Figure 10). Strong system-selection evidence would require several competitive adaptive systems with crossing service–change tradeoffs; the present service-floor analysis should therefore be read only as a boundary result.

Simulation enables attribution but limits generalization. MovieLens supplies initial users, ratings, and movie genres, but the preference dynamics, endorsement rules, service utility, and benchmark-defined costs are simulated. The three response mechanisms span qualitatively different dynamics but cannot establish real-world prevalence. Realistic self-report noise, missing controlled queries, nonlinear or strategic responses, and new domains would require new trajectories. The no-influence arm used to define hidden counterfactual effects also cannot be reconstructed from ordinary deployment logs; evaluating real users would require a separate randomized and ethically reviewed study.

Information separation prevents direct leakage, not shared causes. The save-before-reveal protocol and field restrictions in Appendix E ensure that the observable evaluator cannot directly read latent state, mechanism labels, or paired-world outcomes before producing its estimate. They do not make the controlled check statistically independent of the hidden score: both are designed to respond to simulated preference movement. The benchmark therefore supports controlled attribution within the simulator, not identification from arbitrary observational logs.

J Provenance and Reproduction

The principal result artifacts are:

- frozen held-out summary: `experiments/paper_inputs/v4-heldout-primary-v07/summary.json`;

- frozen observable estimate and revealed truth: `experiments/paper_inputs/v4-heldout-primary-v07/audit/estimate.json` and `experiments/paper_inputs/v4-heldout-primary-v07/truth/revealed.json`;
- post-result held-out diagnostics: `experiments/runs/wlyu-v05-m17-dml1m-q35-9b-heldout-diagnostics-full/audit/heldout_diagnostics.json`;
- artifact-only visual portfolio manifest: `experiments/runs/wlyu-v07-m18-dml1m-artifact-visual-portfolio-full/visual_manifest.json`;
- deterministic figure renderer, specifications, and exact-image audits: `experiments/runs/wlyu-v07-m18-dml1m-artifact-visual-portfolio-full/paper_assets/render_visual_portfolio.py`, `experiments/runs/wlyu-v07-m18-dml1m-artifact-visual-portfolio-full/paper_assets/specs/`, and `experiments/runs/wlyu-v07-m18-dml1m-artifact-visual-portfolio-full/paper_assets/audits/`;
- exploratory smaller-updater summary: `experiments/paper_inputs/v4-crossmodel-qwen3-4b-v04/summary.json`;
- earlier matched controls: `experiments/analysis/m04v_causal_controls_corrected_v1/report.json` and `experiments/analysis/m04v_causal_controls_corrected_v1/bootstrap.json`;
- earlier hashed candidate pool: `experiments/analysis/m05v_union_bounded_weight075_v1/report.json` and `experiments/analysis/m05v_union_bounded_weight075_v1/bootstrap.json`;
- development ranking statistic: `experiments/milestones/M11/results/m11-dml1m-scripted-v4-statlock.json`;
- separation checks: `experiments/milestones/M11/results/m11-dml1m-scripted-v4-independence.json`;
- timeout-only recovery compatibility record: `experiments/analysis/m11_source_bundle_independence_compatibility_f441_20260925.json`;
- same-setting development test: `experiments/milestones/M13/results/m13-dml1m-scripted-anchor-fixed-dose.json`;
- retrospective component analysis: `experiments/milestones/M14/results/m14-dml1m-scripted-check-components.json`;
- retrospective service-floor sensitivity: `experiments/milestones/M15/results/m15-dml1m-scripted-quality-floor.json`;
- distributed Qwen delivery check: `experiments/milestones/M11/results/m11-dml1m-q35-9b-v4-obx-retry.json`.

The Qwen execution uses image digest `sha256:af7687ed83c4706848cc7cf9251bdda741930694cf5c9c1864ec9559e7fe47a8`. The original execution source bundle hash is `bdefc4a8274ef4e3cd2c697999b7193365ff661ad7cae8f0f1001861461ede22`. The final recovery contract records source bundle `f4411a1befe7644e03b4e7b3afbfb24574eef7508a567c5009e5f711f7e041fd`. This recovery changes only the distributed synchronization timeout; the scientific protocol and original launcher are unchanged, as checked by the compatibility record above. The held-out configuration hash is `c7ca772b5c90f3012f9311a4f72504098c6c7213dad5e0ae7afcfb6776194e70`. The distributed held-out result is bound to both source records and to the configuration and container hashes. The MovieLens archive SHA-256 is `a6898adb50b9ca05aa231689da44c217cb524e7ebd39d264c56e2832f2c54e20`.

K Completed Diagnostics and Open Stress Tests

The completed diagnostics evaluate nominal-dose and fixed-path references, same-dose system ordering, fallback sensitivity, lower assessment frequency, doubled controlled-response noise, leave-

one-influence-level-out stability, and both development and held-out service-floor families. None changes the frozen held-out ranking decision. Important open stress tests remain:

- leave one user-response mechanism out of the six-condition average;
- compare recipient matching with an unmatched random assignment.

Response-model mismatch and strategic responses also remain open stress tests. These analyses should hold hidden complete-loop truth fixed so that measurement fragility is not confused with a change in the causal target.