USER HISTORY REVEALS RESIDUAL PREFERENCE IN RECOMMENDER CANDIDATE DISTRIBUTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

User histories can distinguish preferences among recommendation candidates that receive the same observed feedback, yet these distinctions may remain only partially expressed in a trained model's output. We study this *residual user preference* through history-guided refinement of a frozen candidate distribution under explicit probability-preservation constraints. We introduce BehaviorTie-Bench, a controlled benchmark combining real item catalogs from seven domains with programmatic user trajectories, private candidate-level utilities, and shared catalog features for generation and history scoring. We propose *Selective Mass-Preserving History Projection* (SMPHP), which uses candidate-discriminative history scores to refine the teacher's conditional distribution within an eligible subset without additional policy training. The update preserves the subset's total probability mass and every outside-set probability, with exact recovery of the teacher distribution when intervention is rejected. Under the benchmark's feedback-conditioned protocol, SMPHP improves GRPO's mean expected utility by 5.0% on Test-IID and 5.9% on Test-OOD relative to the frozen teacher. Across four teacher configurations, all 44 evaluated held-out domain–teacher–split cells have positive lower bounds in their pointwise 95% bootstrap intervals for the cluster-mean gain. In pooled diagnostics spanning Validation and both test splits, substituting another user's history reverses the mean gain from $+0.01997$ to -0.01299, connecting the improvement to user-specific historical evidence. These controlled results demonstrate recoverable preference in frozen recommendation distributions and, through comparison with unconstrained updates, reveal the utility cost of preserving designated candidate probabilities. Anonymous artifacts: `https://anonymous.4open.science/r/ICLR_smphp-4C3D/`.

1 INTRODUCTION

User interaction histories provide the basis for personalized recommendation, from sequential models to generative recommenders (Hidasi et al., 2016; Kang & McAuley, 2018; Sun et al., 2019; Geng et al., 2022; Rajput et al., 2023). These models map observed behavior to candidate preferences, and preference optimization further refines their predictions through ranking or reward objectives (Chen et al., 2024; Tan et al., 2025; Shi et al., 2026). At the level of an individual recommendation event, however, feedback often identifies a positive response for only a small subset of candidates (Hu et al., 2008; Tan et al., 2025). Candidates with the same observed reward can still differ in their underlying utility to the user. For example, two items receiving no positive feedback may differ in how closely they match a user's earlier positive and negative interactions. This raises a question about the output of a trained recommender: can the recorded history support a better allocation of probability within its current candidate set?

We study this question through *residual user preference*, defined as user-specific information in the recorded history that can improve candidate-level expected utility relative to a frozen teacher. History-informed reranking already uses past interactions to adjust recommendation lists (Steck, 2018); our focus is the recoverable utility when designated candidate probabilities must be preserved. Specifically, we hold the teacher checkpoint and candidate set fixed, permit redistribution only within an eligible subset, and require every outside-set probability and the subset's total probability mass to remain unchanged.

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.

A central obstacle is that ordinary next-item logs do not reveal the utility of every candidate needed to evaluate such a redistribution. We therefore construct BehaviorTie-Bench, a controlled benchmark combining real item catalogs from seven domains with programmatic user trajectories and candidate-level simulator utilities. Seven constructed mechanism slices and IID/OOD splits make residual preference measurable under different feedback, history, and trajectory conditions. The controlled generator and history scorer share catalog topic features, and some slices use utility gaps and structural-history criteria during construction. The main evaluation is feedback-conditioned, with a unique observed current positive protected from probability changes; private utilities are reserved for construction, validation-time gate selection, and offline evaluation.

Building on this setting, we propose *Selective Mass-Preserving History Projection* (SMPHP), an inference-time layer that requires no additional policy training. The method addresses two coupled requirements: extracting history that differentiates the current candidates and applying that evidence within the allowed intervention region. It scores candidate–history relations using behavioral and recency weights, with a variance-based gate that downweights history events whose similarities are nearly uniform across candidates. A separate intervention gate uses agreement between full and recent histories together with score separation to determine whether to apply an edit. For accepted edits, SMPHP mixes the history and teacher conditionals within the eligible set and restores that set's original probability mass; rejected edits return the teacher distribution exactly. The resulting expected-utility change decomposes into eligible probability mass, intervention strength, and the history conditional's utility advantage, linking the effect of an edit to the quality of its history-derived direction.

We evaluate SMPHP with SFT, GRPO, ReRe, and TAWin teachers, comparing each frozen distribution with its projected counterpart. For GRPO, mean expected utility increases from 0.37944 to 0.39838 on Test-IID and from 0.27854 to 0.29498 on Test-OOD, corresponding to relative improvements of 5.0% and 5.9% on the benchmark utility scale. Across pooled diagnostics spanning Validation and both test splits, substituting another user's history reverses the mean gain from $+0.01997$ to -0.01299, linking the improvement to user-specific historical evidence. Unconstrained history injection attains a larger gain while changing protected probabilities, documenting the utility–preservation tradeoff in the evaluated interventions.

Our contributions are summarized as follows:

- We formulate residual user preference as recoverable candidate-level utility in a frozen recommendation distribution, with an explicit boundary on which probabilities may change.
- We introduce BehaviorTie-Bench for controlled diagnosis across seven domains, combining candidate-level simulator utilities with constructed mechanism slices and IID/OOD evaluation.
- We propose SMPHP with exact mass-preservation and fallback properties. Evaluations across four teacher configurations yield positive gains in all 44 held-out domain–teacher–split cells, while history substitution and unconstrained updates characterize the signal's user specificity and its utility–preservation tradeoff.

2 Related Work

Sequential and generative recommendation. GRU4Rec, Caser, SASRec, and BERT4Rec learn mappings from ordered histories to items Hidasi et al. (2016); Tang & Wang (2018); Kang & McAuley (2018); Sun et al. (2019). P5, TIGER, BIGRec, and D3 study recommendation through alternative representations, generation, or decoding Geng et al. (2022); Rajput et al. (2023); Bao et al. (2025; 2024). Our question starts with an already trained recommender and an already available candidate distribution. SMPHP neither retrieves additional items nor replaces the teacher's sequence model.

Preference optimization. BPR and preference/RL methods optimize ranking or policy parameters during training Rendle et al. (2009); Rafailov et al. (2023); Chen et al. (2024); Gao et al. (2025); Shao et al. (2024); Tan et al. (2025); Shi et al. (2026). Our comparison begins after this optimization: each checkpoint is frozen and only its candidate distribution is edited, so training and inference-time refinement are distinct, potentially complementary effects.

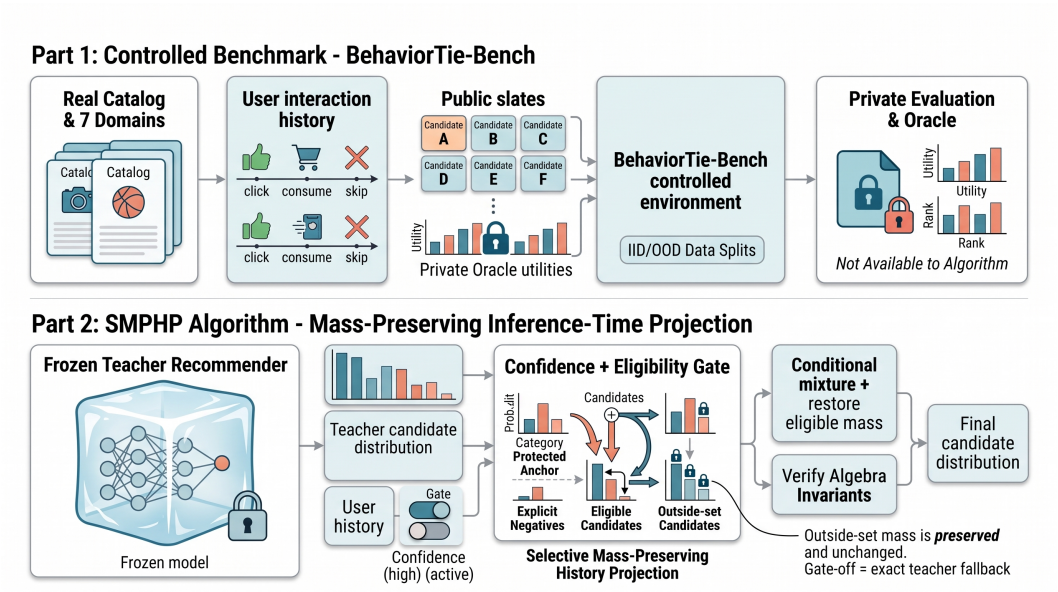

Figure 1: **Controlled benchmark and mass-preserving history projection.** Top: real catalogs support programmatic histories and candidate slates, with private Oracle utilities for controlled evaluation. The Oracle is not a test-time refinement input; controlled gate calibration uses Validation utilities. Bottom: history supplies a candidate preference direction; the teacher supplies the original distribution and eligible-set mass. The conditional mixture changes only eligible probabilities. Rejected intervention returns the teacher, without changing the checkpoint or retrieving new candidates.

History-based post-processing. Calibrated recommendation reranks a model's output to better match the distribution of interests in a user's history Steck (2018). This is a close precedent for using history after scoring. SMPHP instead evaluates candidate-level utility while fixing outside-set probabilities and eligible-set mass; matching a list's category proportions is not its objective. Simple learned history models such as EASE Steck (2019) would motivate stronger comparisons, but neither work is an evaluated baseline here.

Selection and confidence calibration. Selective classification and SelectiveNet trade coverage against prediction risk Geifman & El-Yaniv (2017; 2019); here rejection targets an edit, leaving the teacher's prediction available. Confidence calibration asks whether predicted probabilities reflect correctness frequencies Guo et al. (2017). Our history-agreement score is not such a probability, and its exact property is fallback, not error-rate control.

Logged evaluation and simulation. Propensity-based recommendation evaluation and counterfactual risk minimization address selection bias in logged feedback Schnabel et al. (2016); Swaminathan & Joachims (2015). Randomized-exposure data such as KuaiRand provide a complementary empirical resource Gao et al. (2022). Our benchmark takes a different route: it supplies simulator utilities for every candidate on a fixed slate. RecSim and RecSim NG likewise make user dynamics configurable Ie et al. (2019); Mladenov et al. (2021). BehaviorTie-Bench focuses on frozen-distribution diagnosis across constructed mechanisms, not on introducing a general simulation platform, and simulator utility is not claimed to identify real-world causal outcomes.

3 Problem and Controlled Benchmark

3.1 Observed-Feedback Ambiguity and Residual Preference

At a recommendation event, observed feedback often identifies only a small subset of candidates, leaving several with the same visible outcome. Equal feedback does not imply equal latent utility:

two candidates with the same current response may still differ in how well they match the user's long-term interests, recent state, or session goal. User history can contain candidate-discriminative evidence about these differences even when the frozen model output does not express them.

Let $\mathcal{H}_u = \{(i_t, b_t, x_t)\}_{t=1}^{T}$ be a user's recorded history and let $\mathcal{C} = \{c_1, \ldots, c_K\}$ denote the current candidate slate. A frozen teacher produces

$$p_i^{\mathrm{T}} = P_{\mathrm{T}}(c_i \mid x, \mathcal{H}_u, \mathcal{C}). \tag{1}$$

We use *residual user preference* to denote user-specific information in the recorded history that can support a better candidate-level allocation than this frozen distribution. It is relative to a checkpoint, candidate slate, and evaluation criterion, and does not imply that the teacher ignored history.

Our main question is the *constrained recovery* of this residual. Let $\mathcal{E} \subseteq \mathcal{C}$ be the candidates whose probabilities may change. The admissible intervention set is

$$\mathcal{A}(p^{\mathrm{T}}, \mathcal{E}) = \left\{ p \in \mathbb{R}^K : p_i \geq 0,\ \sum_{i=1}^{K} p_i = 1,\ p_i = p_i^{\mathrm{T}} \text{ for } i \notin \mathcal{E},\ \sum_{i \in \mathcal{E}} p_i = \sum_{i \in \mathcal{E}} p_i^{\mathrm{T}} \right\}. \tag{2}$$

Thus probability can move inside $\mathcal{E}$, while every outside-set probability and the total mass assigned to $\mathcal{E}$ remain fixed. In the controlled benchmark, private candidate utilities U_i define

$$\mathrm{EU}(p) = \sum_{i=1}^{K} p_i U_i, \qquad \Delta\mathrm{EU} = \mathrm{EU}(p^{\mathrm{F}}) - \mathrm{EU}(p^{\mathrm{T}}). \tag{3}$$

A positive $\Delta\mathrm{EU}$ under an admissible history-guided redistribution is our operational evidence that residual preference is recoverable under the specified intervention boundary. EU is expected simulator utility on the fixed slate, not an accuracy percentage or an online conversion rate.

3.2 Controlled Benchmark and Information Boundary

Ordinary next-item logs do not reveal the utility of every candidate needed to evaluate such a redistribution. We therefore use BehaviorTie-Bench as a controlled measurement instrument. It combines real item catalogs from seven domains with programmatic user trajectories and private candidate-level utilities. Each domain contains 800 controlled users, a 1,000-item catalog, 5,000 controlled training events, 2,500 SFT events, and 900 events in each of Validation, Test-IID, and Test-OOD; source counts and simulator dynamics are in Appendix A.

The information boundary is explicit. At inference, the refinement may use the recorded history, candidate metadata, and frozen teacher distribution. Private utilities may be used during benchmark construction and for selecting a gate threshold on Validation; Test utilities are reserved for offline evaluation and never enter the test-time refinement, and public observations and private utilities are stored separately and paired only by sample ID.

The primary controlled runs use L2-normalized 32-dimensional NMF topic features derived from TF–IDF over item titles and categories. These features are shared by the generator and the primary history scorer and are therefore part of the declared benchmark condition. Independent-representation sensitivity is reported in Appendix E.1.

3.3 Mechanism Slices and Evaluation Protocols

The benchmark contains seven constructed slices: Clean Informative, Clean Tie Clear, Clean Tie Ambiguous, Shared Feature Stress, Behavior Flip, Intent Shift, and Logged Noisy. Some slice-selection rules use private utility gaps and structural-history criteria. Slice-level results therefore diagnose mechanisms under selected conditions rather than estimate their natural prevalence. An unfiltered simulator split that removes this mechanism-based selection is reported separately in Appendix E.2.

The main controlled experiment is *feedback-conditioned*: when a unique current external positive is observed, that candidate is protected from probability changes. Controlled thresholds are selected using Validation utilities, while Test utilities are used only for offline evaluation. A separate *pre-feedback* public check excludes current reward and private utility from the refinement and instead protects the teacher's top-1 candidate; because its information set and metrics differ from the main experiment, it is treated only as an external boundary check (Appendix E.5).

4 SELECTIVE MASS-PRESERVING HISTORY PROJECTION

4.1 CANDIDATE-DISCRIMINATIVE HISTORY DIRECTION

The first requirement is to extract from history a direction that distinguishes the current candidates. This differs from retrieving items that are merely similar to the history: a history event is useful here only when its relation to the candidates is non-uniform and can therefore change their relative allocation. For each candidate i and history event t, define

$$M_{it} = \cos(e(c_i), e(i_t)), \quad \mu_t = \frac{1}{K}\sum_i M_{it}, \quad v_t = \mathrm{Var}_i(M_{it}), \quad \gamma_t = \frac{v_t}{v_t + \kappa}. \tag{4}$$

The controlled runs use the shared NMF catalog features described above. With $\kappa = 0.01$, the variance gate γ_t discounts history events that relate similarly to all candidates, suppressing non-discriminative evidence rather than simply favoring the candidates most similar to the history.

Let a_t combine observed behavior/rating/dwell weight with recency, using a 10-day half-life. The de-shared score is

$$\tilde{a}_t = \frac{a_t}{\sum_{t'} |a_{t'}|}, \qquad s_i^{\mathrm{H}} = \sum_t \frac{\tilde{a}_t \gamma_t}{\sum_{t'} |\tilde{a}_{t'} \gamma_{t'}|}(M_{it} - \mu_t). \tag{5}$$

Numerical zero-sum handling is specified in Appendix B. Candidate-wise centering removes the candidate-independent component of each history event; the variance gate changes the relative weight of history events and suppresses nearly uniform evidence. Once the eligible set $\mathcal{E}$ is defined, the scores induce the history conditional

$$q^{\mathrm{H}}_{i|\mathcal{E}} = \frac{\exp(s_i^{\mathrm{H}}/\tau)}{\sum_{j\in\mathcal{E}} \exp(s_j^{\mathrm{H}}/\tau)}, \qquad i \in \mathcal{E}, \quad \tau = 0.1. \tag{6}$$

4.2 INTERVENTION BOUNDARY AND SELECTIVE GATE

The second requirement is to determine where history evidence is allowed to act. Let $\mathcal{P}$ denote protocol-specified protected candidates and $\mathcal{N}$ explicit-negative candidates. The eligible set is

$$\mathcal{E} = \mathcal{C} \setminus (\mathcal{P} \cup \mathcal{N}).$$

In the feedback-conditioned setting, a unique current external positive is protected; in the public pre-feedback setting, the teacher's top-1 candidate is protected. These exclusions define the intervention boundary, not a claim that the excluded candidates have zero utility.

Within this boundary, a heuristic confidence c combines nonnegative Spearman agreement between full- and recent-history scores with entropy-based score separation (Appendix B). A Validation-selected threshold gives $g \in \{0, 1\}$ and

$$\rho = g\rho_{\max} c, \qquad \rho_{\max} = 0.2. \tag{7}$$

The gate decides whether to edit the teacher distribution; it is not a calibrated probability of utility improvement. Histories with fewer than four events, fewer than three unique candidate keys, negligible eligible mass, or fewer than two eligible candidates return the teacher unchanged.

4.3 REFINING THE CANDIDATE DISTRIBUTION

For $M_{\mathcal{E}} = \sum_{i\in\mathcal{E}} p_i^{\mathrm{T}} > 10^{-12}$, let $p^{\mathrm{T}}_{i|\mathcal{E}} = p_i^{\mathrm{T}}/M_{\mathcal{E}}$. We mix the two conditionals inside the eligible set and then restore its original mass:

$$p_i^{\mathrm{F}} = \begin{cases} M_{\mathcal{E}}\left[(1-\rho)p^{\mathrm{T}}_{i|\mathcal{E}} + \rho q^{\mathrm{H}}_{i|\mathcal{E}}\right], & i \in \mathcal{E}, \\ p_i^{\mathrm{T}}, & i \notin \mathcal{E}. \end{cases} \tag{8}$$

The refinement satisfies three exact invariants:

$$p_i^{\mathrm{F}} = p_i^{\mathrm{T}}, \qquad i \notin \mathcal{E}, \tag{9}$$

$$\sum_{i\in\mathcal{E}} p_i^{\mathrm{F}} = \sum_{i\in\mathcal{E}} p_i^{\mathrm{T}}, \tag{10}$$

$$\rho = 0 \Rightarrow p^{\mathrm{F}} = p^{\mathrm{T}}. \tag{11}$$

Protected and explicit-negative candidates inherit the outside-set invariance. These are structural probability guarantees: they do not guarantee rank preservation or positive utility on every accepted edit.

4.4 The Optimization Objective

We now state the objective that the refinement optimizes. Using the same private utility vector on both sides of eq. (8),

$$\Delta\mathrm{EU} = \rho M_{\mathcal{E}} \underbrace{\left[\mathbb{E}_{q_{\mathcal{E}}^{\mathrm{H}}}[U] - \mathbb{E}_{p_{\mathcal{E}}^{\mathrm{T}}}[U]\right]}_{A_{\mathcal{H}}:\ \text{history conditional advantage}} . \tag{12}$$

This identity separates three factors: the teacher mass available to edit $M_{\mathcal{E}}$, intervention strength ρ, and the utility advantage $A_{\mathcal{H}}$ of the history-derived conditional. Because the first two are non-negative, a positive gain requires a useful direction rather than merely a different distribution. It also predicts that wrong-user history should damage the direction, that a weak scorer should reduce gain under the same shell, and that enlarging the intervention region can help only by relaxing the preservation requirement.

5 Experiments

5.1 Experimental Setup and Objectives

The primary comparison is the same frozen teacher with and without SMPHP on identical eight-candidate slates and evaluation utilities. Teachers are LoRA-tuned Qwen2.5-1.5B-Instruct policies Yang et al. (2024), adapted with LoRA Hu et al. (2022); GRPO starts from the domain SFT checkpoint. SMPHP adds no policy-training stage. SFT and GRPO cover seven domains, while ReRe and TAWin cover Industrial, Office, Toys, and Yelp. Each configuration uses training seeds 42/43/44. Unless stated otherwise, the main effectiveness tables use held-out Test-IID and Test-OOD only. Pooled history and refinement diagnostics may include Validation and are labeled as such. Each teacher is compared only with its own frozen baseline; detailed training, aggregation, and interval procedures appear in Appendix C.

The experiments address five objectives: whether residual preference is recoverable at all (section 5.2), whether the recovered direction is specific to the current user (section 5.3), whether the scorer supplies a better direction than simpler alternatives under a matched intervention shell (section 5.4), what the preservation constraint costs in utility (section 5.5), and where the resulting gain does and does not transfer (section 5.6). Aggregation rules and interval definitions are given in Appendix C.

5.2 Residual Preference Is Recoverable

Mean GRPO EU rises from 0.37944 to 0.39838 on Test-IID and from 0.27854 to 0.29498 on Test-OOD (table 1), absolute gains of 0.01895 and 0.01643, or 5.0% and 5.9% relative to the frozen teacher. Every domain improves on both splits, and all 42 GRPO domain–seed–test-split units have positive mean gains. Because the checkpoint and candidate slate are fixed, the improvement comes from redistribution within the existing candidate set rather than retrieval or teacher training, and the OOD macro gain retains about 87% of the IID gain.

The effect extends beyond GRPO. All 44 held-out domain–teacher–split intervals have positive lower bounds: 14 each for SFT and GRPO and eight each for ReRe and TAWin (table 2). The conclusion is teacher-relative: the same constrained refinement improves each frozen teacher on its own held-out scope, and the differing domain coverage prevents ranking teachers.

5.3 The Recovered Direction Is User-Specific

Wrong-user history reverses the gain. The full-history refinement yields a pooled ΔEU of $+0.01997$. Replacing the input with another user's history changes this to -0.01299 while the teacher and candidate slate remain fixed. This sign reversal links the correction to the current user's

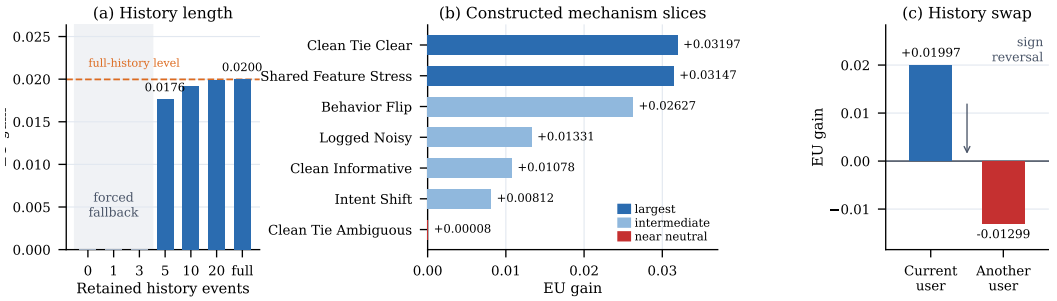

Figure 2: **History length, mechanism slices, and history identity.** (a) Histories below four events force fallback, and five events already recover most of the full-history gain. (b) Slice-level gains under constructed selection; this is a diagnostic, not natural prevalence. (c) Another user's history reverses the pooled gain.

Table 1: **Adding SMPHP improves the frozen GRPO distribution in every reported domain and test split.** Entries are expected utility (EU; higher is better), averaged over three seeds on identical candidate slates. In each split, read left to right: original teacher, the same teacher plus SMPHP, and their paired gain. The last row is the domain macro-average. Gains are read from the result ledger.

	Test-IID			Test-OOD		
Domain	GRPO alone	GRPO + SMPHP	EU gain ↑	GRPO alone	GRPO + SMPHP	EU gain ↑
Beauty	0.40322	0.42157	+0.01835	0.28104	0.29835	+0.01731
Industrial	0.41315	0.43298	+0.01984	0.30991	0.32486	+0.01496
Musical	0.37804	0.39863	+0.02059	0.31321	0.32688	+0.01367
Office	0.40330	0.41943	+0.01612	0.31886	0.33167	+0.01282
Sports	0.39867	0.41893	+0.02026	0.24285	0.26356	+0.02071
Toys	0.41266	0.42738	+0.01472	0.31676	0.33141	+0.01465
Yelp	0.24702	0.26977	+0.02275	0.16719	0.18810	+0.02092
Macro-avg.	0.37944	0.39838	**+0.01895**	0.27854	0.29498	**+0.01643**

Table 2: **Every teacher benefits on its held-out evaluation scope.** Test-IID and Test-OOD only. Each EU gain compares that row's frozen teacher with itself plus SMPHP. The last column counts domain–teacher–split pointwise 95% bootstrap intervals with positive lower bounds. SFT/GRPO cover seven domains and ReRe/TAWin four, so rows are not a matched teacher ranking.

Teacher	Domains	Units	Teacher EU	ΔEU ↑	Positive CIs
GRPO	7	42	0.32899	+0.01769	14/14
SFT	7	42	0.28429	+0.02080	14/14
ReRe	4	24	0.15015	+0.02820	8/8
TAWin	4	24	0.15482	+0.02774	8/8

evidence: an input-independent smoothing operation would not change when the history identity is swapped. The experiment therefore establishes user specificity, not a single scoring component.

A short informative history recovers most of the observed effect. Five events yield $+0.01761$, ten $+0.01915$, and twenty $+0.01995$, against $+0.01997$ for full history; 0/1/3 events force exact fallback by design. Constructed slices localize the effect further: Clean Tie Clear and Shared Feature Stress show the largest gains, while Clean Tie Ambiguous is nearly neutral (fig. 2); detailed perturbations are in Appendix F.

Table 3: **Matched-shell scorer comparison on Test-IID.** Entries are paired EU differences between the listed conditional direction and the full SMPHP scorer, averaged over the available domains for each teacher under the shared NMF representation. Eligibility, protected candidates, gate, intervention strength, and mass constraint are fixed. Values near zero for the full scorer are numerical replay checks, not an additional utility gain; uniform is a direction-free control and the remaining rows remove one component at a time.

Conditional direction	GRPO	SFT	ReRe	TAWin
Full SMPHP	-6.7×10^{-11}	-6.8×10^{-11}	-7.1×10^{-11}	-7.1×10^{-11}
Uniform direction	-0.03082	-0.03093	-0.02990	-0.02981
Signed visible weights	-0.00304	-0.00304	-0.00301	-0.00299
Mean cosine	-0.00282	-0.00282	-0.00243	-0.00242
Without variance gate	-0.00200	-0.00200	-0.00205	-0.00204
Teacher-visible weights	-0.00095	-0.00094	-0.00098	-0.00097
Without dwell weight	-0.00051	-0.00051	-0.00054	-0.00054
Without recency weight	-0.00044	-0.00043	-0.00043	-0.00043
Unit action lookup	-0.00035	-0.00034	-0.00003	-0.00003

Table 4: **Unconstrained gain and constrained intervention are different objectives.** Means over 126 SFT/GRPO units, including Validation. Anchor drift averages absolute per-run mean changes; mass error averages within-run absolute eligible-mass errors. Neither column is a maximum event-level bound.

Method	Only inside $\mathcal{E}$?	ΔEU $\uparrow$	Anchor drift	Mass error
Naive History	No	+0.05456	2.15×10^{-2}	1.21×10^{-2}
Mass-preserving, no gate	Yes	+0.01997	1.37×10^{-18}	5.42×10^{-17}
SMPHP	Yes	+0.01997	1.34×10^{-18}	5.14×10^{-17}

5.4 The Scorer Builds a Better Direction

To separate scorer quality from the intervention boundary, we hold the candidate slate, eligible set, protected candidates, gate decision, intervention strength, and mass-preservation rule fixed and replace only the conditional direction inside $\mathcal{E}$.

Every simpler direction reduces EU relative to the full scorer, and uniform redistribution is an order of magnitude worse than any history-based direction (table 3). The component ordering is stable across the four teachers: the centered, variance-gated construction contributes the most (≈ 0.0020), followed by dwell (≈ 0.0005) and recency (≈ 0.0004) weighting, while the absolute action-lookup scale contributes least and its interval contains zero. Under the shared benchmark representation and the same intervention shell, the candidate-discriminative scorer therefore supplies a more useful direction than these alternatives. This comparison tests the direction term in eq. (12); it is not a claim of superiority to external reranking methods that are not evaluated here.

5.5 The Cost of the Preservation Constraint

Naive History achieves a larger pooled gain ($+0.05456$) but changes protected probabilities and eligible-set mass (table 4). The mass-preserving variants retain a positive gain ($+0.01997$) while satisfying the specified intervention boundary to numerical precision. The result therefore shows an explicit utility–preservation tradeoff rather than an advantage in unconstrained utility maximization. From eq. (8), the update size is also bounded by $\frac{1}{2}\|p^{\mathrm{F}} - p^{\mathrm{T}}\|_1 \leq \rho M_{\mathcal{E}}$, so the intervention is restricted in location and magnitude.

The gate adds an exact fallback decision: on rejected interventions SMPHP returns the teacher distribution, whereas the ungated mass-preserving variant can still move probability. Mean gain is unchanged because fallback only removes rejected edits, so the gate's effect is on the tail rather than the mean (section 5.6). Detailed projection checks are reported in Appendix F.

Table 5: **Gate calibration changes the risk profile, not the mass guarantee.** Controlled Test-IID, shared NMF, equal-weight domain average. Coverage and the Oracle-gate row appear in Appendix E.

	Frozen gate		Recalibrated gate	
Teacher	Active neg.	CVaR@5%	Active neg.	CVaR@5%
GRPO	0.251	-0.0323	0.179	-0.0118
SFT	0.194	-0.0228	0.120	-0.0071
ReRe	0.114	-0.0090	0.075	-0.0036
TAWin	0.114	-0.0099	0.073	-0.0037

Under the same intervention boundary, the mean per-unit Oracle-projection recovery ratio is 72.8% (mean Oracle gain 0.02728 versus 0.01997 for SMPHP). Because both share the eligible set, gate, intervention strength, and mass constraint, the gap isolates direction quality, which remains the main source of the unexploited headroom; it is not the fraction of all user preference recovered. Denominator ranges appear in Appendix D.

5.6 Boundary Checks

Two conditions bound where the controlled gain is available (Appendix E). First, the shared NMF representation is part of the main condition: under independent title encoders the GRPO gain falls from $+0.01895$ to $+0.00116$ (MiniLM) and -0.00094 (BGE-M3), while a residual-information probe still finds mean Spearman correlations of 0.287 and 0.269 against 0.086 for random. Second, an unfiltered simulator split retains about 59% (IID) and 52% (OOD) of the controlled gain on the three domains with a recoverable pool.

Three further checks bound deployment. Under the frozen gate a quarter of accepted GRPO edits lose utility, and Validation recalibration reduces the worst-5% tail from -0.0323 to -0.0118 without lowering mean gain (table 5); a synthetic pre-feedback variant retains 66–73% of the controlled gain; and on public fixed-candidate data reranking changes NDCG@10 only at the 10^{-5}–10^{-4} level and never activates on Yelp.

6 Discussion and Conclusion

Three conclusions follow. First, under the controlled shared-representation protocol a frozen candidate distribution still contains recoverable, user-specific residual preference: redistributing only existing probability mass inside an eligible subset improves expected utility in every evaluated held-out cell, and substituting another user's history reverses the pooled gain. Second, the identity in eq. (12) explains when that information converts into utility; the conversion is representation- and teacher-dependent, and it is roughly halved once mechanism-based selection is removed. Third, mass preservation and fallback are exact structural properties, whereas utility gain, tail-risk reduction, and public transfer are empirical: the frozen gate leaves a quarter of accepted GRPO edits negative, and recalibration reduces that tail without lowering mean gain.

The claim is deliberately narrow: residual preference is relative to a checkpoint, candidate slate, representation, and evaluation utility. We do not claim that the teacher ignores history, that sparse feedback is the only source of residual preference, or that every accepted edit improves utility. BehaviorTie-Bench makes the first question measurable; SMPHP offers one constrained mechanism.

Reproducibility Statement

We release an anonymous artifact containing the benchmark code, catalog features, result tables, and evaluation scripts at `https://anonymous.4open.science/r/ICLR_smphp-4C3D/`. Appendix A describes how each domain and its user trajectories are constructed, Appendix A.2 gives the simulator dynamics, Appendix B states the refinement invariants, and Appendix C reports the training, calibration, and interval settings. The main tables are aggregated from the released result tables following the rules in Appendix C.

AI USE STATEMENT

Generative AI assisted with source inspection, organization, wording, and the preparation of analysis scripts. The authors are responsible for all results, equations, references, and conclusions.

REFERENCES

Keqin Bao, Jizhi Zhang, Yang Zhang, Xinyue Huo, Chong Chen, and Fuli Feng. Decoding matters: Addressing amplification bias and homogeneity issue in recommendations for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 10540–10552, 2024. doi: 10.18653/v1/2024.emnlp-main.589.

Keqin Bao, Jizhi Zhang, Wenjie Wang, Yang Zhang, Zhengyi Yang, Yanchen Luo, Chong Chen, Fuli Feng, and Qi Tian. A bi-step grounding paradigm for large language models in recommendation systems. *ACM Transactions on Recommender Systems*, 3(4), 2025. doi: 10.1145/3716393.

Yuxin Chen, Junfei Tan, An Zhang, Zhengyi Yang, Leheng Sheng, Enzhi Zhang, Xiang Wang, and Tat-Seng Chua. On softmax direct preference optimization for recommendation. In *Advances in Neural Information Processing Systems*, 2024.

Chongming Gao, Shijun Li, Yuan Zhang, Jiawei Chen, Biao Li, Wenqiang Lei, Peng Jiang, and Xiangnan He. KuaiRand: An unbiased sequential recommendation dataset with randomly exposed videos. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management*, pp. 3953–3957, 2022. doi: 10.1145/3511808.3557624.

Chongming Gao, Ruijun Chen, Shuai Yuan, Kexin Huang, Yuanqing Yu, and Xiangnan He. Sprec: Self-play to debias llm-based recommendation. In *Proceedings of the ACM Web Conference 2025*, pp. 5075–5084, 2025. doi: 10.1145/3696410.3714524.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 2151–2159. PMLR, 2019.

Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. Recommendation as language processing (rlp): A unified pretrain, personalized prompt & predict paradigm (p5). In *Proceedings of the 16th ACM Conference on Recommender Systems*, 2022. doi: 10.1145/3523227.3546767.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Ruining He and Julian McAuley. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th International Conference on World Wide Web*, pp. 507–517, 2016. doi: 10.1145/2872427.2883037.

Balazs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. Session-based recommendations with recurrent neural networks. In *International Conference on Learning Representations*, 2016.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Yifan Hu, Yehuda Koren, and Chris Volinsky. Collaborative filtering for implicit feedback datasets. In *2008 Eighth IEEE International Conference on Data Mining*, 2008. URL https://yifanhu.net/PUB/cf.pdf.

Eugene Ie, Chih-Wei Hsu, Martin Mladenov, Vihan Jain, Sanmit Narvekar, Jing Wang, Rui Wu, and Craig Boutilier. RecSim: A configurable simulation platform for recommender systems. *arXiv preprint arXiv:1909.04847*, 2019.

Wang-Cheng Kang and Julian McAuley. Self-attentive sequential recommendation. In *2018 IEEE International Conference on Data Mining*, 2018.

Martin Mladenov, Chih-Wei Hsu, Vihan Jain, Eugene Ie, Christopher Colby, Nicolas Mayoraz, Hubert Pham, Dustin Tran, Ivan Vendrov, and Craig Boutilier. RecSim NG: Toward principled uncertainty modeling for recommender ecosystems. *arXiv preprint arXiv:2103.08057*, 2021.

Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of EMNLP-IJCNLP*, pp. 188–197, 2019. doi: 10.18653/v1/D19-1018.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, 2023.

Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan H. Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Q. Tran, Jonah Samost, Maciej Kula, Ed H. Chi, and Maheswaran Sathiamoorthy. Recommender systems with generative retrieval. In *Advances in Neural Information Processing Systems*, 2023.

Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pp. 452–461, 2009.

Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. Recommendations as treatments: Debiasing learning and evaluation. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1670–1679, 2016. URL https://proceedings.mlr.press/v48/schnabel16.html.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Wentao Shi, Qifan Wang, Chen Chen, Fei Liu, Dongfang Liu, Xu Liu, Wanli Ma, Junfeng Pan, Linhong Zhu, and Fuli Feng. Objective shaping with hard negatives: Windowed partial auc optimization for rl-based llm recommenders. *arXiv preprint arXiv:2604.22504*, 2026.

Harald Steck. Calibrated recommendations. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pp. 154–162, 2018. doi: 10.1145/3240323.3240372.

Harald Steck. Embarrassingly shallow autoencoders for sparse data. In *The World Wide Web Conference*, 2019. doi: 10.1145/3308558.3313710.

Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pp. 1441–1450, 2019. doi: 10.1145/3357384.3357895.

Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 814–823, 2015. URL https://proceedings.mlr.press/v37/swaminathan15.html.

Junfei Tan, Yuxin Chen, An Zhang, Junguang Jiang, Bin Liu, Ziru Xu, Han Zhu, Jian Xu, Bo Zheng, and Xiang Wang. Reinforced preference optimization for recommendation. *arXiv preprint arXiv:2510.12211*, 2025.

Jiaxi Tang and Ke Wang. Personalized top-n sequential recommendation via convolutional sequence embedding. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pp. 565–573, 2018. doi: 10.1145/3159652.3159656.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

A Dataset Construction and State Dynamics

A.1 Source Catalogs and Benchmark Scale

The six product domains use Amazon review and item-metadata resources He & McAuley (2016); Ni et al. (2019); Yelp supplies the seventh catalog. These resources ground item identities and descriptions, not the simulated users' private utilities.

For each source domain, the construction pipeline (1) builds a real-item catalog with nonempty titles, (2) derives normalized topic features from title/category TF–IDF using NMF, (3) simulates longitudinal profiles and 50-step trajectories, (4) samples eight-item slates, (5) generates public history/actions and private candidate utilities, and (6) exports Public and Private records matched by `sample_id`. Split assignment precedes mechanism construction, but some slice-selection rules still use Oracle and structural-history criteria. Record separation does not remove this construction dependence.

Table 6: Constructed benchmark scale. Source rows are raw domain interactions; users and training/evaluation events are controlled. Each evaluation-count entry lists Validation/Test-IID/Test-OOD.

Domain	Source rows	Sim. users	Items	Train	Val/IID/OOD
Beauty	198,502	800	1,000	5,000	900/900/900
Industrial	45,324	800	1,000	5,000	900/900/900
Musical	199,532	800	1,000	5,000	900/900/900
Office	48,656	800	1,000	5,000	900/900/900
Sports	296,337	800	1,000	5,000	900/900/900
Toys	140,943	800	1,000	5,000	900/900/900
Yelp	96,961	800	1,000	5,000	900/900/900

Across the seven domains, every catalog entry has a nonempty title, every evaluation slate contains eight distinct candidates, public observations and private utilities are stored separately and joined only by sample identifier, and the projection invariants hold to machine precision. The source rows sum to 1,026,255; the generated catalogs contain 7,000 items, and the controlled training sets contain 35,000 events. Controlled timestamps are simulator-relative seconds, not source-calendar timestamps. Counts do not estimate real-world prevalence.

A.2 Generator and State Dynamics

The simulator generates each user trajectory from the following dynamics. Let f_i be the item's topic-feature vector, p_u, n_u the long-term positive and negative profiles, r_t, z_t the recent and recent-negative states, and g_t the session goal. The state-dependent preference components are

$$L_i = f_i^\top p_u, \quad R_i = f_i^\top r_t, \quad G_i = Q_i = f_i^\top g_t, \quad N_i = f_i^\top (n_u + z_t). \tag{13}$$

For 50 steps indexed by $t = 0, \ldots, 49$, gradual drift uses $g_t = (1 - t/49)p_u + (t/49)p'_u$; abrupt shift switches from p_u to p'_u at step 25; multi-interest alternates by step parity; ambiguous-sparse uses their equal mixture; stable and rejection-sensitive profiles use p_u. OOD overrides these schedules with triangular interpolation: $\phi_t = (t \bmod 10)/9$, replaced by $1 - \phi_t$ in odd ten-step blocks, and $g_t = (1 - \phi_t)p_u + \phi_t p'_u$.

Given the eight candidate utilities, the sampled item follows

$$P(i \mid u, t) = (1 - \eta_u)\,\mathrm{softmax}(U/T_{\mathrm{choice}})_i + \eta_u/8, \qquad \eta_u = \min(0.15\,\mathrm{exploration}_u, 0.50). \tag{14}$$

Conditional on its utility, the action distribution is $P(a \mid U) = \mathrm{softmax}((m_a U + b_a)/T_{\mathrm{action}})$.

Starting from zero states, with persistence $\lambda_u \sim \mathrm{Uniform}(0.78, 0.94)$, the selected item updates

$$r_{t+1} = \lambda_u r_t + (1 - \lambda_u)\xi_a f_i, \tag{15}$$

$$z_{t+1} = 0.90 z_t + 0.10\,\mathbf{1}[\xi_a < 0] f_i. \tag{16}$$

Table 7: Simulator action parameters. These update latent states and are distinct from the deployed scorer's behavior weights.

Action	m_a	b_a	ξ_a
skip	-1.2	0.1	0.0
click	-0.4	0.5	0.2
long_view	0.1	0.1	0.4
save	0.35	-0.1	0.6
consume	0.75	-0.5	1.0
rate_positive	1.0	-0.8	1.0
rate_negative	-0.7	-0.2	-0.5
reject	-1.1	-0.5	-1.0

The action is appended after the event's history snapshot. External reward in the logged simulator is one for the selected item only for long-view, save, consume, or positive rating; other entries are zero. Clean/tie construction subsequently changes feedback and is not identical to this logged process.

Low/medium/high noise tuples $(T_{\text{choice}}, T_{\text{action}}, \sigma_\epsilon)$ are $(0.35, 0.45, 0.03)$, $(0.65, 0.70, 0.06)$, and $(1.00, 1.00, 0.12)$. Train/Validation/IID use approximately equal proportions; OOD uses 30% medium and 70% high noise. Users are assigned to splits before slice construction. Base histories are capped at twenty events; Shared Feature Stress may append four nodes. Near-equality of twenty-event and full-history results is relative to this bounded protocol.

A.3 From a Source Item to a Constructed Evaluation Record

The Beauty catalog contains 12,087 source items, of which 1,000 are selected. TF–IDF over titles and categories yields 2,274 features and 32 topics, with no zero-feature rows. The simulator uses 800 users and 50-step trajectories. Its long-term, recent, and goal coefficients are 0.50, 0.25, and 0.25; negative, query, novelty, repetition, and popularity coefficients are 1.10, 0.35, 0.20, 0.35, and 0.10. These are configured simulator parameters, not independently estimated human preferences.

The slice configuration sets minimum history length four, clear and ambiguous Oracle-gap thresholds 0.10 and 0.05, and clear structural pair accuracy 0.70. Shared Feature Stress adds four consume events, prioritizing low candidate-relation variance and then high mean relation. Behavior Flip changes one positive action to reject. These selection criteria differ from the Validation-selected deployment gate. Five unpaired slices contribute 100 records each and two paired slices contribute 200 each, giving 900 records per domain/evaluation split. The index contains 49 representative domain–slice groups, not a sample of natural mechanism frequencies.

Appendix G.1 shows one concrete Beauty record: its full 13-event history, the exact changed action in its paired variant, and all eight candidate utilities and probabilities. The generator and Oracle are programmatic, whereas the recommendation teacher is learned. The controlled scorer uses the shared catalog features; no external text encoder is used in these evaluation units.

Source record and stored representation. The case spans 17 distinct items across its history and candidates. For example, its metadata excerpt identifies `B0014UAHQW`, "NOW Foods Rose Hip Seed Oil, 1 ounce," at line 47,706 of `meta_Beauty.json.gz`, with brand "Now Foods" and category path Beauty / Skin Care / Body / Moisturizers / Oils. The catalog retains this identity and title. Its stored 32-dimensional vector has leading components 0.763622, 0.643135, and 0.037539 at topic indices 24, 8, and 18, respectively. These are the stored catalog feature values.

A concrete simulated step. The corresponding source event is `controlled_beauty_u_00458:0013`, after 13 recorded interactions. Its logged exposure paths comprise five random candidates, two popularity candidates, and one recent-similarity candidate: the rose hip oil. For that item the saved weighted utility contributions are

$$\begin{aligned} U_{\text{oil}} = {} & 0.630144_{\text{long}} + 0.168818_{\text{recent}} + 0.315072_{\text{goal}} + 0.441101_{\text{query}} \\ & - 0.021553_{\text{negative}} + 0.022885_{\text{novelty}} - 0.350000_{\text{repetition}} \\ & - 0.068789_{\text{popularity}} + 0.042661_{\text{noise}} = 1.180339. \end{aligned}$$

The source event records selection of this item at zero-based position 1, with selection probability 0.750791 and action `rate_positive`. The stored recent-state component at topic 24 changes from 0.427870 to 0.500189 after this step. This is a programmatic user-state update, not an observed change in a real person's preference.

From the source event to the evaluated slice. The source and final records contain the same item–utility pairs, but *not the same ordering or feedback*. The following mapping uses the A–H candidate labels of Appendix G.1; source indices are zero-based. It prevents accidentally joining utilities by row position.

Final label	Item ID	Source exposure	Source index	Saved U
A	B003FP86SK	random	2	0.032156
B	B007BLN17K	popularity	3	-0.110916
C	B0017UEU62	random	0	-0.018496
D	B0014UAHQW	recent similarity	1	1.180339
E	B000ZMBSPE	popularity	5	0.518885
F	B001ANOONK	random	6	-0.616272
G	B00188IFHS	random	7	-0.741925
H	B001720JL0	random	4	0.002823

The source reward vector is $(0, 1, 0, 0, 0, 0, 0, 0)$; the final positive-member slice has eight zero external rewards. Consequently, the source event's selected item must not be treated as a protected current positive in the evaluated slice. Within the final Behavior Flip pair, candidates, their order, and utilities do remain fixed: history index 10 changes from castor-oil `consume` to `reject`, with corresponding polarity, prompt, and identity changes. For the positive member, the frozen SFT probability on the rose hip oil changes from 0.034109 to 0.222101 under SMPHP, producing the $+0.265209$ event EU gain detailed in Appendix G.

What this trace shows. The construction record joins the 17 source-metadata excerpts and feature rows to the history and slate, matches the 13 preceding actions to the final history, and matches all eight source utilities to the final utilities by item ID. Summing the nine stored components for each candidate reproduces its stored utility within 1.01×10^{-6}, the precision at which the components are exported. The case is therefore inspectable record by record; it is not a bit-exact replay from the latent state, and it does not by itself establish real-human validity.

A.4 An Annotated Training Record: All Twenty History Events

The preceding example connects construction to a frozen evaluation result. Here we inspect a different object: a serialized *training* record, `v1_2:pilot_5k_clean_informative:train:1821:observed`, for simulated user `controlled_beauty_u_00635` and source event `controlled_beauty_u_00635:0025`. The record comes from the same archived Beauty run as the sample-order list. It is a training example, not an evaluated test case and not evidence of a positive projection gain; it illustrates what a training example contains and how its target is represented.

Identity, context, and time. The stored fields are `split=train`, `mechanism_slice=pilot_5k_clean_informative`, `query_condition=partial_query`, `logging_policy=popularity_heavy`, and `behavior_noise=medium`. The current request is "Recommend beauty products related to topic 24 oil argan." The target timestamp is 2,160,000 seconds (simulated day 25). All twenty history timestamps precede it, spanning days 5–24; day here means the stored timestamp divided by 86,400, not a calendar date.

Complete retained history. The table lists every history event in the record, in order. Titles are shortened for readability; item IDs preserve identity. Action codes, rating fields, and dwell values are retained. A zero rating is the stored field value, not an inferred zero-star judgment.

No.	Day	Item ID	Short title	Action	Rating	Dwell
1	5	B003K9KZHG	Sunny Isle castor oil	save	0	0.362181
2	6	B000JVA80C	Earth Mama stretch oil	click	0	0.178214
3	7	B000ZMBSPE	Aztec bentonite clay	skip	0	0.007755
4	8	B002WTC37A	Macadamia repair mask	consume	0	0.870437
5	9	B00JJVG6HC	Dead Sea mud shampoo	rate_negative	2	0.138597
6	10	B0069FDR96	World Pride nail tape	skip	0	0.046953
7	11	B000HRVC5I	Beautyblender sponge	click	0	0.118517
8	12	B00GXBEV0A	Morocco argan hair mask	save	0	0.425733
9	13	B001JHZ6W0	Bulgarian rose water	click	0	0.245129
10	14	B000JVA80C	Earth Mama stretch oil	save	0	0.415398
11	15	B00150LT40	Seche Vite top coat	click	0	0.155341
12	16	B003K9KZHG	Sunny Isle castor oil	long_view	0	0.606357
13	17	B00JJVG6HC	Dead Sea mud shampoo	long_view	0	0.534351
14	18	B002WTC37A	Macadamia repair mask	click	0	0.208070
15	19	B008V0U33U	Hair One almond cleanser	rate_negative	2	0.198281
16	20	B000JVA80C	Earth Mama stretch oil	click	0	0.224409
17	21	B003V265QW	Remington curling wand	reject	0	0.029260
18	22	B003K9KZHG	Sunny Isle castor oil	save	0	0.344470
19	23	B00KD73PBQ	Argan hair oil serum	long_view	0	0.524728
20	24	B003I7IB62	Luscious facial oil	consume	0	0.982971

The record contains four saves, six clicks, two consumes, three long views, two negative ratings, one rejection, and two skips. Its stored polarities classify fifteen events as positive, three as negative, and two as uncertain. The shampoo has both a negative rating and a later long view; collapsing this history to a set of item IDs would discard that distinction. The clay's skip is labeled uncertain rather than negative. These are concrete reasons to retain action and temporal information, not proof that the projection improves this particular training example.

A.5 Eight Candidates and the Provenance of the Training Label

The same record exposes exactly eight candidates. The table preserves their final A–H order and the full reward vector; titles are shortened. The last column joins candidates to the history above by item ID, not by title.

Choice	Item ID	Short title	Reward	Earlier history in this record
A	B000ZMBSPE	Aztec bentonite clay	0	day 7: skip
B	B0043OYFKU	Olay Pro-X cleansing system	0	no exact item-ID match
C	B002WTC37A	Macadamia repair mask	0	day 8: consume; day 18: click
D	B00639DLV2	Hawaiian Tropic SPF 30 sunscreen	1	no exact item-ID match
E	B004D24818	Neutrogena eye repair	0	no exact item-ID match
F	B00JJVG6HC	Dead Sea mud shampoo	0	day 9: rate_negative; day 17: long_view
G	B0052G15UG	Nail-polish display rack	0	no exact item-ID match
H	B004OHQR1Q	Nail-art dotting pens	0	no exact item-ID match

The label can be checked directly. The stored vector is $r = (0, 0, 0, 1, 0, 0, 0, 0)$, so the unique reward argmax is zero-based position 3, corresponding to choice D. This agrees with both `sft_target_action=D` and the final assistant message `D`. In contrast, `selected_action=save` names a behavior, not a choice letter. The sample also stores `choice_permutation=[1,5,4,7,6,3,2,0]`, a permutation of 0–7 applied to the source candidate array.

Two stages of label provenance must not be conflated. The record declares

```
reward_source: v1.2_private_oracle_top1_controlled
sft_label_source:
public_external_reward_unique_argmax;private_oracle_not_used
```

Read together, these fields describe the declared lineage

$$\text{Oracle-derived controlled reward} \longrightarrow \text{stored public reward vector} \longrightarrow \text{unique-argmax SFT label D.}$$

The downstream label-extraction step can avoid reading a private file while still inheriting an Oracle-derived training signal. Consequently, `private_oracle_not_used` must not be read as "Oracle-independent dataset construction." Nor is `reward_reliability=1.0` an empirical estimate of agreement with real users. The record declares `choice_protocol=clean_verifier` and `reward_tie=false`; these describe the controlled labeling protocol.

What this example does not show. The sunscreen's lack of an exact history match does not make its label an error: titles and observed history alone do not determine the private utility vector. This record therefore shows the public-reward argmax rather than the underlying Oracle optimum. It also contains no matched teacher or final probabilities, so no EU, gain, gate activation, or probability preservation can be reported from it. Under the feedback-conditioned anchor rule, D would be protected; that is a protocol statement, not a projection outcome. In particular, the shampoo's mixed history does not by itself establish the implementation's final explicit-negative mask.

A.6 Stored Messages Are Not All Inference Inputs

The serialized record contains structured fields, a user `prompt`, and a `messages` list with system, user, and assistant roles. Selected strings from the stored messages are shown below. The history and candidate lists are represented by the complete tables in the preceding two subsections; they are not omitted from the original user message.

> **System:** You are a recommender system. Return exactly one choice letter from A to H and no explanation.
>
> **User message structure:** Observable user behavior history: [the twenty numbered events, including the two 2/5 ratings].
>
> **Current request:** Recommend beauty products related to topic 24 oil argan.
>
> **Exposed candidate choices:** [the eight A–H candidates above].
>
> **Closing instruction:** Choose one exposed candidate. Only output its choice letter.
>
> **Stored assistant target:** D

Field-level information boundary. The table distinguishes semantic roles in this record. It specifies how an evaluation reader must separate context from supervision; it does not certify the behavior of an uninspected data loader.

Stored information	Role and required separation
Query, candidate titles/order, history actions/titles/ratings	Recommendation context rendered in the user message, with ratings shown when nonzero. Item IDs remain structured alignment keys, not literal strings in this prompt.
History timestamps and dwell values	Structured history information. Present in the record but not printed as numeric fields in this user message; availability to a separate history scorer must be checked at its interface.
Assistant message `D`; `sft_target_action`	Training target, not observed context. It must not be supplied as a completed answer when generating a test prediction.
`external_rewards`; `selected_action`	Current feedback and behavior. May serve the declared training or feedback-conditioned protocol, not a pre-feedback prediction input.
`reward_source`; `sft_label_source`; mechanism/split/identity fields	Provenance and grouping metadata, not additional evidence available to a pre-feedback recommender.
Private utilities, teacher/final probabilities, gate and eligible mask	Not stored in this record. They cannot be reconstructed from the assistant target or the one-hot reward vector.

For SFT, the assistant completion is the intended supervision. For evaluation, the observed prefix and candidate set must be separated from that completion; candidate likelihood scoring can score hypothetical A–H completions without providing the known correct answer as observed context. The existence of an assistant message inside an exported training record is therefore not alone evidence of test leakage. Conversely, calling a file "Public" does not make every stored field a valid pre-feedback input.

Consistency checks. The stored arrays have matching lengths (twenty history entries and eight candidates), candidate IDs are unique, timestamps increase and precede the target, and the reward argmax agrees with the serialized target. Together with the construction trace and the evaluation cases, this example exposes the recorded data lineage, the training representation, and the measured intervention, without treating those three objects as interchangeable.

B Refinement Details and Invariants

B.1 Score Normalization and Confidence

In eq. (5), the first absolute-sum denominator is clipped below at 10^{-12}; if the gated absolute sum in the second denominator is below 10^{-12}, the score vector is set to zero. The controlled configuration

uses catalog TF-IDF/NMF features. Although the generic scorer supports sentence encoders, that capability is not evidence that the formal runs used them.

Confidence is the product of full/recent-history agreement and score separation:

$$c = \max(0, \mathrm{Spearman}(s^{\mathrm{H}}, s^{\mathrm{recent}})) \left(1 - \frac{H(\mathrm{softmax}(s^{\mathrm{H}}/\tau))}{\log K_{\mathrm{unique}}}\right). \tag{17}$$

Recent scores use the latter half of history. Scores are averaged across duplicate candidate keys before confidence is computed. Fewer than four history events or fewer than three unique keys force $c = 0$. Confidence is a heuristic; the controlled threshold is selected on Validation Oracle utility rather than interpreted as a probability that the modification improves utility.

B.2 Mass Preservation and Fallback

Both eligible conditionals sum to one. Hence

$$\sum_{i\in\mathcal{E}} p_i^{\mathrm{F}} = M_{\mathcal{E}} \left[(1-\rho)\sum_{i\in\mathcal{E}} p_{i|\mathcal{E}}^{\mathrm{T}} + \rho \sum_{i\in\mathcal{E}} q_{i|\mathcal{E}}^{\mathrm{H}}\right] \tag{18}$$

$$= M_{\mathcal{E}}. \tag{19}$$

Outside-set probabilities equal the teacher values by eq. (8). Consequently the full distribution remains normalized, and protected candidates and explicit negatives excluded from $\mathcal{E}$ are unchanged. If $\rho = 0$, the inside-set expression reduces to $M_{\mathcal{E}} p_{i|\mathcal{E}}^{\mathrm{T}} = p_i^{\mathrm{T}}$. With negligible eligible mass or fewer than two eligible candidates, the implementation returns the teacher rather than defining an invalid conditional.

B.3 Utility Decomposition

Outside-set contributions cancel when subtracting expected utilities. For $i \in \mathcal{E}$, the probability difference is $\rho M_{\mathcal{E}}(q_{i|\mathcal{E}}^{\mathrm{H}} - p_{i|\mathcal{E}}^{\mathrm{T}})$. Multiplication by U_i and summation gives eq. (12). This identity explains possible gains and losses but does not guarantee that the history conditional has greater utility. Constant scores imply a uniform history conditional, not equality with the teacher; exact fallback follows from the intervention rule, not from uniformity alone.

C Training, Calibration, and Statistics

C.1 Training and Calibration Settings

For each controlled domain, teacher, and training seed, gate calibration considers all distinct structurally eligible Validation confidence values and the next representable value below their minimum. It seeks coverage at least 0.20 and active negative-gain rate at most 0.35, maximizing all-event mean utility gain; ties favor lower risk, then a higher threshold. The code relaxes the risk constraint if necessary and records this fact; this fallback is not a risk guarantee. Across the 42 domain–teacher–seed calibration records, the selected thresholds range from 0.014332 to 0.166080, and the largest observed active negative-gain fraction is 0.209418. Training seeds 42/43/44 are independent training runs on a single simulated benchmark, not independently regenerated benchmarks.

Controlled SFT uses LoRA on all linear layers with rank 16, alpha 32, dropout 0.05, float16, maximum length 1152, learning rate 10^{-4}, one epoch, per-device batch size 8, and gradient accumulation 2. GRPO starts from the domain SFT checkpoint, uses the same LoRA configuration, learning rate 5×10^{-6}, maximum completion length 4, eight generations, generation batch size 8, per-device batch size 8, $\beta = 0.04$, temperature 0.9, top-$p = 0.9$, and 5,000 optimizer steps with dataset shuffling disabled. Formal seeds are 42/43/44.

SMPHP uses $\rho_{\max} = 0.2$, a 10-day recency half-life, de-shared scoring with $\kappa = 0.01$, and history temperature 0.1. Controlled thresholds are calibrated on Validation only. Formal runs used NVIDIA Tesla V100 32GB GPUs. The public evaluation uses no private Oracle utility and no current reward as input, and neither the gate, the eligible set, ρ, nor the history score uses the target. Candidate alignment checks report zero invalid and zero duplicated candidate actions, with maximum projection mass error 4.44×10^{-16}.

C.2 Metrics, Aggregation, and Interval Interpretation

The primary outcome is the paired expected-utility difference in eq. (3). Oracle probability is probability on the Oracle top candidate. For a fixed slate and utility vector, absolute expected regret would be $\max_i U_i - \mathrm{EU}(p)$; its reduction equals $\Delta\mathrm{EU}$ under identical aggregation. It is therefore not an independent source of improvement evidence.

Formal intervals use 20,000 mechanism-unit bootstrap resamples with seed 20260808. Event gains are averaged within each mechanism unit across three training seeds, and unit means are resampled. Equal weighting of unit means can differ from equal weighting of individual events, as made explicit in table 8. Intervals are pointwise 95% intervals conditional on the benchmark and frozen training runs. They do not automatically account for dependence between distinct mechanism units from the same user or source trajectory, and are not simultaneous intervals over all cells. Validation intervals are descriptive because that split also selects the gate.

D Additional Controlled Results

D.1 Oracle Projection Recovery

Under the same eligible set, gate, ρ, and mass constraints, let $\Delta\mathrm{EU}_n^{\mathrm{OracleProj}}$ be the gain from an Oracle-directed conditional for evaluation unit n. The reported recovery is

$$\text{CM-OPR} = \frac{1}{N}\sum_{n=1}^{N}\frac{\Delta\mathrm{EU}_n^{\mathrm{SMPHP}}}{\Delta\mathrm{EU}_n^{\mathrm{OracleProj}}}, \qquad \Delta\mathrm{EU}_n^{\mathrm{OracleProj}} > 0, \tag{20}$$

the arithmetic mean of 126 per-domain/teacher/seed/split ratios, not the ratio of pooled gains. It includes Validation and has no separate ratio confidence interval. Evaluation-unit Oracle gains range from 0.020409 to 0.033877, so none of the reported ratios has a zero or near-zero denominator. The comparison uses the same eligible set, gate, intervention strength, and mass constraints; it is not a global Oracle bound.

Ratios range from 61.2% to 79.7% across units. Domain means range from 70.2% (Office) to 75.6% (Yelp). The mean ratio is 74.4% for SFT, 71.2% for GRPO, 74.3% on Test-IID, and 68.9% on Test-OOD. Across all 126 units, the mean SMPHP gain is 0.01997 EU and 0.01665 Oracle probability; the mean Oracle projection gain is 0.02728. These pooled descriptions do not replace the held-out, split-specific comparison in the main table.

D.2 How Much Does Adding SMPHP Improve Each Teacher?

Table 8 compares **each teacher plus SMPHP against that same teacher alone**: SFT, GRPO, ReRe, and TAWin. It does not compare teachers against one another. For each row, the checkpoint and candidate slate are fixed; SMPHP is the added inference layer. All gain columns use

$$\text{EU gain} = \mathrm{EU}(\text{teacher + SMPHP}) - \mathrm{EU}(\text{teacher alone}).$$

Thus $+0.01835$ for Beauty/GRPO/Test-IID means that adding SMPHP improves EU by 0.01835, not that either model's absolute EU is 0.01835. The corresponding before/after scores are 0.40322 and 0.42157 in table 1.

Two averages, not two models. The event-mean and cluster-mean columns summarize the *same paired gain* using different weights. The event mean weights events equally; the cluster mean weights mechanism-unit means equally. The 95% interval belongs only to the cluster mean. For the example above, the event-mean gain is $+0.01835$ and the cluster-mean gain is $+0.01582$, with interval $[0.01354, 0.01831]$. These are not the teacher's score and SMPHP's score. Absolute SFT before/after scores are not reported in this gain ledger and cannot be inferred from it.

The 66 rows comprise 44 held-out Test-IID/Test-OOD rows and 22 Validation rows. SFT/GRPO contribute 42 rows across seven domains; ReRe/TAWin contribute 24 across Industrial, Office, Toys, and Yelp. All held-out lower bounds are positive. Validation rows are descriptive after gate selection, not additional held-out confirmation. Three-seed comparisons are aggregated before intervals are formed.

Table 8: **EU improvement from adding SMPHP to each frozen teacher.** The comparison column names the enhanced model and its own baseline. Positive gains favor the model with SMPHP; they are absolute utility changes, not relative percentages. Event and cluster columns are two averages of the same gain, not before/after model scores. Pointwise 95% intervals describe cluster gains only and are not multiplicity-adjusted. The 66 rows separate 44 held-out intervals from 22 descriptive Validation intervals.

Domain	Comparison	Split	EU gain ↑ Event mean	EU gain ↑ Cluster mean	95% CI for cluster gain
Beauty	GRPO + SMPHP vs. GRPO alone	Test-IID	+0.01835	+0.01582	[0.01354, 0.01831]
		Test-OOD	+0.01731	+0.01541	[0.01360, 0.01732]
		Validation	+0.01846	+0.01604	[0.01398, 0.01819]
	SFT + SMPHP vs. SFT alone	Test-IID	+0.02222	+0.01925	[0.01702, 0.02169]
		Test-OOD	+0.02023	+0.01819	[0.01637, 0.02010]
		Validation	+0.02300	+0.01980	[0.01761, 0.02210]
Industrial	GRPO + SMPHP vs. GRPO alone	Test-IID	+0.01984	+0.01712	[0.01508, 0.01923]
		Test-OOD	+0.01496	+0.01313	[0.01140, 0.01496]
		Validation	+0.01961	+0.01693	[0.01488, 0.01904]
	SFT + SMPHP vs. SFT alone	Test-IID	+0.02314	+0.02010	[0.01803, 0.02227]
		Test-OOD	+0.01771	+0.01561	[0.01382, 0.01747]
		Validation	+0.02340	+0.02013	[0.01798, 0.02236]
Musical	GRPO + SMPHP vs. GRPO alone	Test-IID	+0.02059	+0.01786	[0.01570, 0.02010]
		Test-OOD	+0.01367	+0.01231	[0.01056, 0.01412]
		Validation	+0.01775	+0.01541	[0.01335, 0.01757]
	SFT + SMPHP vs. SFT alone	Test-IID	+0.02347	+0.02049	[0.01830, 0.02277]
		Test-OOD	+0.01606	+0.01445	[0.01270, 0.01625]
		Validation	+0.02138	+0.01850	[0.01636, 0.02074]
Office	GRPO + SMPHP vs. GRPO alone	Test-IID	+0.01612	+0.01394	[0.01212, 0.01583]
		Test-OOD	+0.01282	+0.01124	[0.00972, 0.01284]
		Validation	+0.01927	+0.01672	[0.01469, 0.01885]
	SFT + SMPHP vs. SFT alone	Test-IID	+0.02134	+0.01853	[0.01648, 0.02066]
		Test-OOD	+0.01670	+0.01477	[0.01313, 0.01648]
		Validation	+0.02433	+0.02118	[0.01906, 0.02337]
Sports	GRPO + SMPHP vs. GRPO alone	Test-IID	+0.02026	+0.01784	[0.01576, 0.01998]
		Test-OOD	+0.02071	+0.01815	[0.01615, 0.02025]
		Validation	+0.01933	+0.01672	[0.01478, 0.01874]
	SFT + SMPHP vs. SFT alone	Test-IID	+0.02363	+0.02073	[0.01860, 0.02294]
		Test-OOD	+0.02256	+0.01977	[0.01778, 0.02186]
		Validation	+0.02326	+0.02012	[0.01811, 0.02220]
Toys	GRPO + SMPHP vs. GRPO alone	Test-IID	+0.01472	+0.01289	[0.01129, 0.01458]
		Test-OOD	+0.01465	+0.01284	[0.01107, 0.01473]
		Validation	+0.01640	+0.01377	[0.01206, 0.01551]
	SFT + SMPHP vs. SFT alone	Test-IID	+0.01955	+0.01708	[0.01534, 0.01896]
		Test-OOD	+0.01795	+0.01577	[0.01397, 0.01766]
		Validation	+0.02188	+0.01854	[0.01652, 0.02058]
Yelp	GRPO + SMPHP vs. GRPO alone	Test-IID	+0.02275	+0.01969	[0.01761, 0.02179]
		Test-OOD	+0.02092	+0.01829	[0.01631, 0.02031]
		Validation	+0.02504	+0.02227	[0.02023, 0.02443]
	SFT + SMPHP vs. SFT alone	Test-IID	+0.02474	+0.02152	[0.01938, 0.02371]
		Test-OOD	+0.02191	+0.01928	[0.01729, 0.02130]
		Validation	+0.02677	+0.02385	[0.02177, 0.02605]

Continued on the next page.

Table 8 continued: gain = EU(teacher + SMPHP) − EU(teacher alone).

Domain	Comparison	Split	EU gain ↑ Event mean	EU gain ↑ Cluster mean	95% CI for cluster gain
Industrial	ReRe + SMPHP vs. ReRe alone	Test-IID	+0.03759	+0.03335	[0.03054, 0.03629]
		Test-OOD	+0.02610	+0.02337	[0.02105, 0.02578]
		Validation	+0.03855	+0.03348	[0.03062, 0.03643]
	TAWin + SMPHP vs. TAWin alone	Test-IID	+0.03611	+0.03203	[0.02930, 0.03489]
		Test-OOD	+0.02572	+0.02305	[0.02082, 0.02538]
		Validation	+0.03764	+0.03263	[0.02985, 0.03548]
Office	ReRe + SMPHP vs. ReRe alone	Test-IID	+0.03264	+0.02868	[0.02603, 0.03142]
		Test-OOD	+0.02541	+0.02280	[0.02061, 0.02502]
		Validation	+0.03789	+0.03318	[0.03031, 0.03610]
	TAWin + SMPHP vs. TAWin alone	Test-IID	+0.03179	+0.02795	[0.02535, 0.03065]
		Test-OOD	+0.02506	+0.02257	[0.02038, 0.02479]
		Validation	+0.03735	+0.03267	[0.02981, 0.03563]
Toys	ReRe + SMPHP vs. ReRe alone	Test-IID	+0.02836	+0.02501	[0.02269, 0.02745]
		Test-OOD	+0.02443	+0.02139	[0.01930, 0.02357]
		Validation	+0.03280	+0.02819	[0.02551, 0.03093]
	TAWin + SMPHP vs. TAWin alone	Test-IID	+0.02706	+0.02376	[0.02148, 0.02618]
		Test-OOD	+0.02372	+0.02075	[0.01870, 0.02289]
		Validation	+0.03148	+0.02708	[0.02449, 0.02974]
Yelp	ReRe + SMPHP vs. ReRe alone	Test-IID	+0.02793	+0.02447	[0.02211, 0.02684]
		Test-OOD	+0.02318	+0.02057	[0.01850, 0.02274]
		Validation	+0.03028	+0.02724	[0.02495, 0.02961]
	TAWin + SMPHP vs. TAWin alone	Test-IID	+0.02864	+0.02514	[0.02272, 0.02757]
		Test-OOD	+0.02381	+0.02116	[0.01902, 0.02343]
		Validation	+0.03095	+0.02777	[0.02545, 0.03017]

D.3 Supplementary Teacher Summaries

Table 9: Secondary teacher summaries. Oracle probability is probability on the Oracle top candidate. Held-out columns exclude Validation; pooled columns include it. Pooled unit counts are 63 each for SFT/GRPO and 36 each for ReRe/TAWin. Every gain compares a teacher with its own projection.

Teacher	Held-out Δ Oracle prob.	Pooled ΔEU	Pooled Δ Oracle prob.
GRPO	+0.01493	+0.01826	+0.01538
SFT	+0.01730	+0.02168	+0.01793
ReRe	+0.02157	+0.03043	+0.02283
TAWin	+0.02135	+0.02994	+0.02259

The positive held-out changes in Oracle probability complement the EU gains: the projection increases mean probability on the best-utility candidate as well as the utility-weighted average over candidates. These diagnostics use the same events and private utilities, so they are complementary outcomes rather than independent replications.

E Boundary and Robustness Analyses

E.1 Where Residual Information Becomes Utility

The matched-shell result identifies the value of the scorer under the shared catalog representation. We next separate two questions that are often conflated: whether a representation contains candidate-level residual information, and whether that information is converted into end-to-end EU improvement by the projection. Table 10 reports the first-stage conversion under the controlled frozen shell

Table 10: **Representation sensitivity under the controlled frozen shell.** Entries are Test-IID EU gains relative to each frozen teacher, averaged over the available domains. The shared NMF representation is the primary controlled condition; independent title encoders and random vectors are sensitivity controls.

Representation	GRPO	SFT	ReRe	TAWin
Shared NMF	$+0.01895$	$+0.02259$	$+0.03163$	$+0.03090$
Title MiniLM	$+0.00116$	$+0.00475$	$+0.01572$	$+0.01506$
Title BGE-M3	-0.00094	$+0.00264$	$+0.01318$	$+0.01253$
Random control	-0.00658	-0.00303	$+0.00755$	$+0.00691$

on Test-IID. The shared 32-dimensional NMF representation is the primary condition; title-only MiniLM and BGE-M3 representations are independent sensitivity conditions, and random vectors are an information-poor control.

The conversion is representation- and teacher-dependent. For GRPO, replacing shared NMF with either independent title representation reduces the mean gain to approximately zero; for SFT, MiniLM retains a small positive point estimate, so the representation effect is not uniformly a sign reversal. ReRe and TAWin retain larger gains under independent title representations, while their random controls remain positive, indicating that part of the observed movement can also arise from redistribution within the eligible set. These comparisons therefore qualify the main result rather than establish representation-independent generalization.

To test whether the loss of end-to-end gain means that residual information has disappeared, we separately compute a residual-information probe without the gate, intervention strength, or projection. On the eligible candidates, the history scores are compared with the private candidate utilities. Across 56,700 controlled events per representation, the mean Spearman correlations are 0.558 for NMF, 0.287 for MiniLM, 0.269 for BGE-M3, and 0.086 for random; the corresponding positive-correlation event rates are 0.902, 0.750, 0.734, and 0.576. Thus, independent title representations still carry candidate-level information above the random control, but that information is not guaranteed to produce a positive EU change after gating, mixing, and mass-preserving projection. In terms of eq. (12), the probe measures whether a direction can be related to utility, whereas the end-to-end result tests whether that direction yields a positive $A_{\mathcal{H}}$ at the available $M_{\mathcal{E}}$ and ρ.

E.2 Unfiltered Natural Split Retains Partial Gain

The controlled mechanism slices use Oracle gaps and structural-history criteria during construction, so their results diagnose mechanisms under selected conditions rather than estimating natural-frequency effects. To test whether the gain survives without slice-based selection, we evaluate an unfiltered natural split drawn uniformly from the available simulator generation pool, without filtering by Oracle gap, structural pair accuracy, scorer gain, or confidence. This split is available only for domains with a recoverable simulator pool and is reported separately from the seven-domain controlled matrix.

Table 11: **Unfiltered natural split retains partial controlled gain.** GRPO only, covering Beauty, Sports, and Yelp. The natural split is drawn uniformly without mechanism-based selection. Controlled comparison uses the same three domains. Natural gains are approximately 59% (IID) and 52% (OOD) of their matched controlled counterparts.

Condition	Test-IID ΔEU	Test-OOD ΔEU
Natural split (Beauty/Sports/Yelp)	$+0.01212$	$+0.01030$
Controlled (same three domains)	$+0.02045$	$+0.01965$
Retention ratio	59.3%	52.4%

The natural-split macro-average gains are $+0.01212$ on Test-IID and $+0.01030$ on Test-OOD (table 11). These remain positive, but they retain approximately 59% and 52% of the controlled gain when compared against the same three domains. User-block bootstrap intervals for the natural split

have positive lower bounds in both splits. This attenuation shows that mechanism-based slice selection amplifies the observed effect, but does not eliminate the residual entirely. The natural split does not estimate real-world prevalence; it tests whether the gain persists when Oracle/structural filtering is removed from the benchmark construction.

E.3 Gate Calibration and Risk

Mass preservation and fallback are exact structural properties; whether the gate also reduces downside risk is a separate empirical question. We compare the frozen protocol gate with a gate recalibrated on Validation for each representation and teacher, and with an Oracle gate that activates only on positive-gain events. All entries use the shared NMF representation and controlled Test-IID events, averaged with equal weight over the available domains.

Table 12: **Gate calibration changes the risk profile, not the mass guarantee.** Controlled Test-IID, shared NMF representation, equal-weight domain average. Coverage is the fraction of events left active; active negative rate is the fraction of active edits with negative EU change; CVaR@5% is the mean of the worst 5% of event changes. The final row is an Oracle gate that activates only on positive-gain events and is not deployable.

	Frozen protocol gate			Validation-recalibrated gate		
Teacher	Coverage	Active neg.	CVaR@5%	Coverage	Active neg.	CVaR@5%
GRPO	0.871	0.251	-0.0323	0.773	0.179	-0.0118
SFT	0.892	0.194	-0.0228	0.803	0.120	-0.0071
ReRe	0.957	0.114	-0.0090	0.851	0.075	-0.0036
TAWin	0.957	0.114	-0.0099	0.849	0.073	-0.0037
Oracle gate	0.672	0	0	—	—	—

Under the frozen gate, a quarter of the accepted GRPO edits lose utility and the worst 5% of events average -0.0323 EU (table 12); the worst single-event loss in the pooled tail is -0.1651. Recalibrating on Validation cuts the active negative rate from 0.251 to 0.179 and CVaR@5% from -0.0323 to -0.0118 while mean gain is essentially unchanged, and the same pattern holds for the other three teachers. An Oracle gate adds only 0.0008 mean gain for GRPO at coverage 0.77, so recalibration already captures most of the available risk reduction. A Validation cross-fit harm predictor reaches AUROC 0.736, 0.759, 0.792, and 0.780 under shared NMF and 0.62–0.68 under independent representations: moderate predictability, not a risk guarantee. The supported statement is that the structural guarantee is fallback, and that risk filtering requires recalibration rather than being inherited from the frozen threshold. Pooled risk numbers in the released files must be read for full-scorer runs only, because the same tables also contain simple-scorer variants.

The gain is not confined to near-ties. Evaluated at the level of a deterministic serving decision, top-1 corrections outnumber harms by $8.8\times$, $8.6\times$, $10.8\times$, and $8.0\times$ across GRPO, SFT, ReRe, and TAWin, with mean argmax utility gains of $+0.0254$, $+0.0407$, $+0.0847$, and $+0.0881$ (frozen shell, shared NMF, Test-IID). Harm rates remain non-zero, consistent with the accepted-edit losses above and with the statement that no individual accepted edit is guaranteed to improve utility.

E.4 Pre-Feedback Retention

Finally we ask how much of the effect survives when the current reward and the protected current positive are unavailable, which is the serving-time situation. In a synthetic pre-feedback variant that keeps the same slates and changes only the protection mask, the gain retains 69% (GRPO), 73% (SFT), 67% (ReRe), and 66% (TAWin) of the corresponding controlled gain. The intervention strength is still taken from the controlled protocol and is not re-tuned here, so this is a boundary-sensitivity diagnostic rather than a deployment result.

E.5 Public Reranking Provides a Separate Pre-Feedback Check

The public protocol freezes ReRe candidate sets and excludes the current target, reward, private utilities, and mechanism labels from the projection. Mean NDCG@10 changes are $+4.96 \times 10^{-5}$ in Industrial, $+1.29 \times 10^{-4}$ in Office, and $+1.50 \times 10^{-4}$ in Toys; MRR also increases in these domains. Yelp returns the teacher exactly because the gate never activates. HR@10 is unchanged. Appendix H reports all values and gate activity.

These observations show how the layer behaves on public fixed candidates without current feedback: small positive mean ranking changes where it acts, and exact fallback in Yelp. Without paired intervals they do not establish a reliable external ranking improvement. The metrics and anchor policy differ from controlled EU evaluation, so their magnitudes cannot be directly compared or their differences attributed to the anchor alone. Retrieval and online impact require separate evaluation.

F History Diagnostics and Refinement Invariants

F.1 History Variants

Table 13: History diagnostics averaged over 126 SFT/GRPO units, including Validation. Gains are relative to the corresponding frozen teacher. Events 0/1/3 force fallback; Uniform Action is a compound intervention.

History variant	ΔEU ↑	History variant	ΔEU ↑
Events 0	0.00000	Events 5	0.01761
Events 1	0.00000	Events 10	0.01915
Events 3	0.00000	Events 20	0.01995
Full history	0.01997	History swap	-0.01299
Shuffle actions	0.01436	Shuffle timestamps	0.01848
Window 3d	0.00079	Window 7d	0.01987
Window 14d	0.02002	Uniform Action	0.02183

F.2 Detailed Refinement Checks

Table 14: Projection-structure checks over 126 SFT/GRPO units, including Validation. Drift and gate-off columns average absolute values of per-run mean changes, not per-event absolute errors. Mass error averages within-run absolute eligible-mass errors. These distinct aggregation rules must not be conflated.

Method	ΔEU	Anchor drift	Outside-$\mathcal{E}$ drift	Mass error	Gate-off $\lvert\Delta\text{EU}\rvert$
Naive History	0.05456	2.15×10^{-2}	5.51×10^{-3}	1.21×10^{-2}	1.44×10^{-2}
Mass-preserving, no gate	0.01997	1.37×10^{-18}	5.31×10^{-19}	5.42×10^{-17}	1.24×10^{-4}
SMPHP	0.01997	1.34×10^{-18}	5.20×10^{-19}	5.14×10^{-17}	2.74×10^{-18}

Across 126 formal units, the maxima of absolute run-mean anchor, explicit-negative, and outside-set changes are 5.51×10^{-18}, 1.17×10^{-18}, and 2.18×10^{-18}. The largest run-mean eligible-mass error is 5.93×10^{-17}; gate-rejected ΔEU is 1.79×10^{-17}. These are structural checks, not universal utility guarantees or per-event maximum-error bounds. A mass-preserving reranker cannot recover an item absent from the candidate pool.

G Actual Histories, Candidate Probabilities, and Case Outcomes

G.1 A Concrete Evaluation Case

The success case is `v1_2:behavior_flip:test_iid:0014:positive`, from Beauty, SFT, seed 42, Test-IID. Its source record is `controlled_beauty_u_00458:0013`. The tables below expose all 13 history events and all eight candidates, rather than only reporting the final EU. The items are real; the user trajectory and utilities are generated by the simulator.

Table 15: The complete visible history of the Beauty success case. Day is the recorded timestamp divided by 86,400; the decision occurs on day 13. Titles are shortened for display. The positive rating for rose hip oil is on day 5. The paired reject variant changes the day-10 consume action, not the candidate utilities.

Day	Item ID	Item (short title)	Action
0	B0055FF3SE	Life-Flo apricot oil	long_view
1	B000ELP5KA	NOW tea tree oil	consume
2	B007XA49O8	NOW avocado oil	consume
3	B001E15III	MONOI TIARE coconut oil	rate_positive
4	B005AO9SLS	Elemis frangipani body oil	consume
5	B0014UAHQW	NOW rose hip seed oil	rate_positive
6	B0055FF3SE	Life-Flo apricot oil	long_view
7	B001LFM4WA	Aveeno shower/bath oil	rate_positive
8	B00A0J09GG	Axe face hydrator/post-shave gel	save
9	B007XA49O8	NOW avocado oil	save
10	B0013TM9UQ	NOW castor oil	consume
11	B005AO9SLS	Elemis frangipani body oil	click
12	B002BPY7H4	Doo Gro growth oil	rate_positive

What changes during slice construction? In the paired record ending in `0014:reject`, the day-10 interaction with castor oil changes from `consume` to `reject`. History polarity and the rendered prompt change accordingly, together with the variant identity. The eight candidate IDs, their order, and their complete utility vector stay fixed. This is an observable record-level example of Behavior Flip: it tests response to an altered history observation, not a real user's causal utility change. The positive member is the success example analyzed here.

What is protected in this particular event? All current external rewards are zero, all eight candidates are eligible, and there is no anchor or explicit-negative candidate. Thus $M_{\mathcal{E}} = 1$ in this example. It illustrates recovery from history, but cannot by itself demonstrate anchor protection. The Industrial failure below supplies an example with explicit-negative candidates and $M_{\mathcal{E}} < 1$.

G.2 The Eight-Candidate Probability Change

The rose hip oil appears in the history with a positive rating, yet SFT assigns it probability 0.034109. SMPHP raises this to 0.222101. Beyond the aggregate gain, the candidate-level contributions identify where utility is recovered and where it is lost in this event.

Table 16: All eight candidates of the same Beauty event, in exported order. $\Delta_i = (p_i^{\mathrm{F}} - p_i^{\mathrm{T}})U_i$. Values are rounded after calculation; full titles and item IDs are preserved in the source records.

ID	Item (short title)	p^{T}	p^{F}	U_i	Δ_i
A	Exfoliating crystals	0.053661	0.043225	0.032156	-0.000336
B	Nail art brushes	0.068902	0.055497	-0.110916	+0.001487
C	Blemish concealer	0.113600	0.091493	-0.018496	+0.000409
D	Rose hip seed oil	0.034109	0.222101	1.180339	+0.221894
E	Bentonite clay	0.170533	0.137342	0.518885	-0.017223
F	Orange blossom lip balm	0.390355	0.314353	-0.616272	+0.046838
G	Lavender conditioner	0.084420	0.067995	-0.741925	+0.012186
H	Eau de toilette	0.084420	0.067994	0.002823	-0.000046

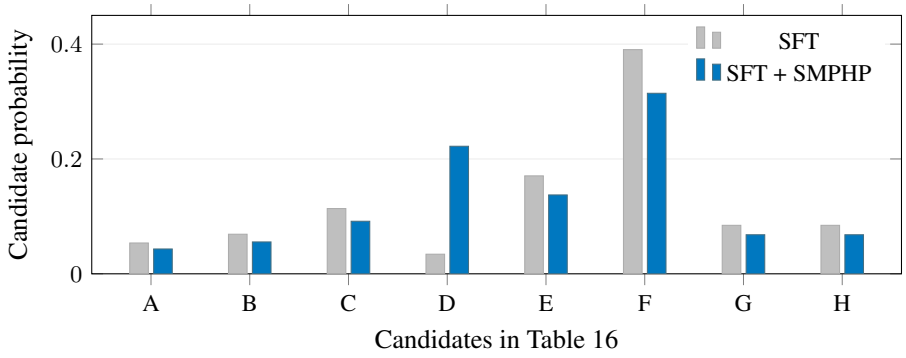

Figure 3: The actual probability change for this success case. Candidate D receives more mass; the candidate set and teacher are unchanged. The table reports the corresponding candidate-level utility contributions.

Why EU improves. The probability increase for rose hip oil contributes $+0.221894$; reducing mass on the negative-utility lip balm contributes $+0.046838$. Reducing mass on positive-utility clay instead costs 0.017223. Summing all eight contributions gives $\Delta\mathrm{EU} = +0.265209$, moving EU from -0.182231 to 0.082979 with $\rho = 0.194734$. The lip balm remains top-ranked, so this example demonstrates improved expected utility, not a corrected top-1 choice.

G.3 Failures and a Real Gate-Off Event

Table 17: SFT cases, seed 42, Test-IID. S1/L1/L2 are the original three examples. W1/W2 minimize gain in the two 900-event units, not in the entire benchmark. R1 is a real gate-off event.

Case	Domain/slice	Gate	ρ	Base EU	+SMPHP EU	ΔEU
S1	Beauty / Behavior Flip	on	0.1947	-0.1822	0.0830	+0.2652
L1	Beauty / Tie Ambiguous	on	0.0128	0.8316	0.8305	-0.0011
L2	Industrial / Tie Ambiguous	on	0.0065	0.9083	0.9058	-0.0026
W1	Beauty / Behavior Flip	on	0.1050	0.7346	0.6462	-0.0884
W2	Industrial / Logged Noisy	on	0.1190	-0.0302	-0.0621	-0.0320
R1	Beauty / Behavior Flip	off	0	0.4005	0.4005	0

A valid edit can still lose utility. In L2, PVC tape with utility 1.357432 loses probability from 0.550436 to 0.547932, contributing -0.003399, partly offset by other candidates. Six candidates are eligible, with mass 0.914802. Two explicit-negative candidates remain at probabilities 0.061757 and 0.023440. Preservation of these probabilities does not guarantee a good direction within the editable set. W1 and W2 lose 0.088412 and 0.031953, showing that the original small-loss examples alone do not characterize the downside.

A real fallback, not a fabricated short history. R1 has 20 history events and is structurally eligible, but its confidence is zero, below the Beauty threshold 0.048279. With $\rho = 0$, its EU remains 0.400548. The Beauty and Industrial units contain 105 and 30 gate-off events, respectively, and no histories shorter than four. Short-history fallback is therefore a protocol rule, not an observed case in this subset.

Identity and selection. All IDs start with `v1_2:<slice>:test_iid:`. S1, W1, and R1 use `behavior_flip` with suffixes `0014:positive`, `0000:reject`, and `0009:reject`; L1 and L2 use `clean_tie_ambiguous` with `0093:observed` and `0027:observed`; W2 uses `logged_noisy:0019:observed`. R1 is the lexicographically first structurally eligible gate-off sample in the Beauty unit. These post-hoc cases explain behavior, not outcome frequency.

G.4 Case-Level Recomputation

We recompute these two evaluation units event by event. Candidate identifiers and private utilities align, teacher probabilities reproduce the softmax of the saved log-likelihoods, and after eligible-set normalization the reconstructed active probabilities match the saved vectors within 3.34×10^{-16}. Gate-off rows use the teacher, and their EU matches the saved gated-method EU within 5.56×10^{-16}.

The advantage in eq. (12) uses only eligible candidates in both expectations, and it remains valid when $M_{\mathcal{E}} < 1$; substituting the full teacher EU for the eligible contribution would be an error. Appendix A.3 traces the corresponding Beauty item from its source metadata to its stored utility components.

H Public Evaluation and Retrieval Context

H.1 Fixed-Candidate Reranking

This comparison freezes ReRe's candidate sets and teacher. The three-seed mean ranking changes are positive in Industrial, Office, and Toys; Yelp returns the teacher exactly. Without paired intervals, these changes are directional evidence only. Fixed-set reranking cannot change HR@10 when the entire ten-candidate set is evaluated.

Table 18: Public fixed-candidate changes, three-seed means. Absolute metric differences for the same ReRe teacher before and after projection; no paired confidence intervals are available.

Domain	ΔNDCG@10	ΔMRR	ΔHR@10	Gate active
Industrial	$+4.96 \times 10^{-5}$	$+5.54 \times 10^{-5}$	0	0.979
Office	$+1.29 \times 10^{-4}$	$+1.51 \times 10^{-4}$	0	0.994
Toys	$+1.50 \times 10^{-4}$	$+1.86 \times 10^{-4}$	0	0.994
Yelp	0	0	0	0

H.2 End-to-End Retrieval Context

This is a different comparison: none of the methods below runs SMPHP. The table contextualizes retrieval, rather than measuring the projection's incremental effect. Side-by-side domain blocks are for layout only; rows do not define cross-domain method pairs.

Table 19: Three-seed mean public retrieval metrics. The methods have different retrieval/generation mechanisms; these are not direct SMPHP comparisons.

	Industrial			Office	
Method	HR@10	NDCG@10	Method	HR@10	NDCG@10
Caser	0.09949	0.06597	Caser	0.10851	0.07235
GRU	0.09368	0.06031	BIGRec	0.09172	0.07949
BIGRec	0.09295	0.07541	D3	0.06864	0.05810
S-DPO	0.07368	0.03466	S-DPO	0.07234	0.03360
GRPO	0.00471	0.00328	GRPO	0.00897	0.00490
TAWin	0.00228	0.00139	TAWin	0.01493	0.00874
	Toys			Yelp	
Method	HR@10	NDCG@10	Method	HR@10	NDCG@10
D3	0.05375	0.03708	GRU	0.04823	0.02390
BIGRec	0.04843	0.03694	S-DPO	0.04348	0.01998
Caser	0.03070	0.01849	Caser	0.03895	0.01949
S-DPO	0.02857	0.01288	D3	0.02942	0.01614
GRPO	0.00374	0.00196	GRPO	0.00327	0.00164
TAWin	0.00267	0.00130	TAWin	0.00254	0.00111

The GRPO/TAWin checkpoints retrieve less effectively than the listed classical baselines. This limits their end-to-end performance, but does not explain the ReRe changes in table 18. Controlled EU and public ranking also use different metrics and information conditions. A within-slate edit cannot retrieve a missing target; neither table establishes online benefit.

I Complete Trace of One Case: From a Source Record to a Measured Gain

This appendix follows a single evaluation case through the whole pipeline, from the raw product record to the final utility number, so that every intermediate object can be inspected in one place. Appendices A.3, G.1, and G give the construction chain and the case tables in more detail; here we only concatenate them.

Stage 1: the raw product record. The seven catalogs are built from Amazon review and item-metadata resources and from Yelp. For Beauty, 12,087 source items are reduced to a catalog of 1,000 items with nonempty titles. One catalog entry is `B0014UAHQW`, "NOW Foods Rose Hip Seed Oil, 1 ounce," recorded at line 47,706 of `meta_Beauty.json.gz`, with brand "Now Foods" and category path Beauty / Skin Care / Body / Moisturizers / Oils.

Stage 2: the catalog representation. TF–IDF over titles and categories yields 2,274 features, which NMF reduces to 32 topics. The stored 32-dimensional vector of this item has leading components 0.763622, 0.643135, and 0.037539 at topic indices 24, 8, and 18. These features are shared by the generator and by the history scorer, and are therefore part of the declared benchmark condition (Section 3).

Stage 3: the simulated user and event. The simulated user `controlled_beauty_u_00458` carries a long-term positive profile p_u, a negative profile n_u, a recent state r_t, a recent-negative state z_t, and a session goal g_t. At source event `controlled_beauty_u_00458:0013`, after 13 recorded interactions, the exposed slate contains five random candidates, two popularity candidates, and one recent-similarity candidate: the rose hip oil. The simulator selects that item with probability 0.750791 and records the action `rate_positive`. The recent-state component at topic 24 moves from 0.427870 to 0.500189 as a result. This is a programmatic state update, not an observed change in a real person's preference.

Stage 4: candidate utilities for the slate. Each candidate carries a stored utility assembled from nine components. For the rose hip oil,

$$\begin{aligned} U_{\text{oil}} = {} & 0.630144_{\text{long}} + 0.168818_{\text{recent}} + 0.315072_{\text{goal}} + 0.441101_{\text{query}} \\ & - 0.021553_{\text{negative}} + 0.022885_{\text{novelty}} - 0.350000_{\text{repetition}} \\ & - 0.068789_{\text{popularity}} + 0.042661_{\text{noise}} = 1.180339. \end{aligned}$$

The eight utilities of the final slate are listed in Table 20; they are the same item–utility pairs as in the source event.

Stage 5: from the logged event to the evaluated slice. The final records keep the same item–utility pairs but change ordering and feedback, so utilities must be joined by item ID rather than by row position. This matters for one reason: the source reward vector is $(0, 1, 0, 0, 0, 0, 0, 0)$, whereas the final positive-member slice has eight zero external rewards. The source event's selected item therefore must *not* be treated as the protected current positive in the evaluated slice. The paired Behavior Flip variant changes one history action — at history index 10, castor-oil `consume` becomes `reject` — while the eight candidate IDs, their order, and their utilities stay fixed.

Stage 6: the first model output. The frozen SFT teacher reads the 13-event history and the eight exposed candidates, and returns the probability vector p^{T} in Table 20. Despite the rose hip oil appearing in the history with a positive rating on day 5, the teacher assigns it only $p^{\mathrm{T}}_D = 0.034109$.

Stage 7: the refinement. History scores are formed from candidate–history similarity, behavioral and recency weights, candidate-wise centering, and the variance gate (Section 4). The intervention gate accepts this edit. The intervention strength is $\rho = 0.194734$, and because all eight candidates are eligible, with no protected anchor and no explicit negative, $M_{\mathcal{E}} = 1$ in this example.

Stage 8: the final computed result. Inside the eligible set the refinement mixes the teacher and history conditionals and restores the set's original mass, yielding p^{F} in Table 20. Multiplying the

probability change of each candidate by its own utility gives the per-candidate contributions $\Delta_i = (p_i^{\mathrm{F}} - p_i^{\mathrm{T}})U_i$. The rose hip oil alone contributes $+0.221894$, and reducing mass on the negative-utility lip balm contributes $+0.046838$; reducing mass on the positive-utility clay costs 0.017223. Summing the eight contributions gives

$$\Delta\mathrm{EU} = +0.265209, \qquad \mathrm{EU}(p^{\mathrm{T}}) = -0.182231 \;\longrightarrow\; \mathrm{EU}(p^{\mathrm{F}}) = 0.082979.$$

The lip balm remains the top-ranked candidate, so this case shows improved expected utility rather than a corrected top-1 prediction.

Table 20: Complete trace of one evaluation case (`v1_2:behavior_flip:test_iid:0014:positive`, Beauty, SFT, seed 42, Test-IID). Utilities are joined by item ID. $\Delta_i = (p_i^{\mathrm{F}} - p_i^{\mathrm{T}})U_i$ and the eight contributions sum to $\Delta\mathrm{EU} = +0.265209$ with $\rho = 0.194734$ and $M_{\mathcal{E}} = 1$.

ID	Item (short title)	p^{T}	p^{F}	U_i	Δ_i
A	Exfoliating crystals	0.053661	0.043225	0.032156	−0.000336
B	Nail art brushes	0.068902	0.055497	−0.110916	+0.001487
C	Blemish concealer	0.113600	0.091493	−0.018496	+0.000409
D	Rose hip seed oil	0.034109	0.222101	1.180339	+0.221894
E	Bentonite clay	0.170533	0.137342	0.518885	−0.017223
F	Orange blossom lip balm	0.390355	0.314353	−0.616272	+0.046838
G	Lavender conditioner	0.084420	0.067995	−0.741925	+0.012186
H	Eau de toilette	0.084420	0.067994	0.002823	−0.000046

Table 21: The same case as a chain. Every stage is inspectable in the released records; later stages are not re-derived from earlier ones by bit-exact replay.

Stage	Object	Value
Raw record	Beauty catalog entry, metadata line 47,706	`B0014UAHQW`, "NOW Foods Rose Hip Seed Oil, 1 ounce"
Representation	32-topic NMF vector from title/category TF–IDF	0.763622 / 0.643135 / 0.037539 at topics 24 / 8 / 18
Simulated event	`controlled_beauty_u_00458:0013`	selection probability 0.750791, action `rate_positive`
Candidate utility	nine stored components for the rose hip oil	1.180339
Evaluated slice	eight candidates, utilities joined by item ID	source reward $(0, 1, 0, 0, 0, 0, 0, 0)$; eight zero final rewards
Teacher output	frozen SFT probability on the rose hip oil	0.034109
Refinement	accepted edit, $\rho = 0.194734$, $M_{\mathcal{E}} = 1$	$p_D^{\mathrm{F}} = 0.222101$
Final result	paired expected-utility change	$\Delta\mathrm{EU} = +0.265209$; EU $-0.182231 \to 0.082979$

What this trace does and does not show. The items are real; the user trajectory, the actions, and the utilities are generated by the simulator. The stored components reproduce the stored utility within 1.01×10^{-6}, the precision at which they are exported, so the case is inspectable record by record. It is not a bit-exact replay from the latent state, and it does not by itself establish real-human validity. The single $+0.265209$ event is one hand-checked case: the pooled mean over 126 units is $+0.01997$, and the failure cases in Appendix G show losses of the same order, so this trace illustrates the mechanism rather than the outcome frequency.