RAISE: DIAGNOSING ACQUISITION COLLAPSE IN COSTLY LLM SIGNALS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) are increasingly used as costly, on-demand components in real systems, but calling them indiscriminately can waste substantial compute, latency, and serving budget. The key deployment question is therefore not only whether an LLM helps on average, but when it is worth calling. We identify a common failure mode, which we call *acquisition collapse*: an LLM signal can appear useful in aggregate or post hoc, yet still provide too little before-call information to support reliable selective use. We introduce **RAISE** (Reward-SNR Actionability in Signal Evaluation), a pre-routing diagnostic framework for testing whether a costly signal is truly actionable before committing to a routing strategy. We instantiate RAISE with Structured Hypothesis Embeddings (SHE), a frozen-LLM intent signal for recommendation using one LLM call per user, and evaluate it through controlled, retrospective, and fresh-cohort studies and a prospective pilot whose audit decisions are frozen before independent outcomes are revealed. Across these settings, predictable incremental benefit, not average lift alone, is what determines whether selective LLM calling becomes deployable, while seemingly strong oracle or subgroup gains can disappear under independent evaluation. More broadly, RAISE reframes costly inference as an information-acquisition problem: before paying for an expensive model, tool, sensor, or measurement, first test whether its value is predictable at decision time. This principle motivates cost-aware acquisition in settings ranging from agent tool use and stronger-model consultation to robotic sensing and clinical decision pipelines.

1 INTRODUCTION

Large language models (LLMs) are increasingly used as optional components inside larger AI systems: recommenders can call an LLM to infer user intent, agents can invoke tools or stronger models, and predictive systems can acquire additional semantic information when needed (Yao et al., 2023; Xi et al., 2024; Ren et al., 2024; Wei et al., 2024). These capabilities come with nontrivial compute, latency, and serving costs. Recent work therefore studies how to route queries across models or selectively invoke expensive components under a limited budget (Chen et al., 2024; Ong et al., 2025; Dekoninck et al., 2025). At deployment scale, the question is no longer simply **whether an LLM helps, but which calls are worth paying for**.

However, evidence that a costly signal is useful does not necessarily show that it can be used selectively. Average-use evaluation asks whether performance improves when the signal is acquired broadly. Routing asks a harder question: **can information available before the call predict where that improvement will occur?** A signal may produce positive average lift, or appear highly valuable for examples selected after their outcomes are observed, while still offering too little predictable structure to support a reliable acquisition policy. Noise and post-selection can further exaggerate apparent subgroup value (Cawley & Talbot, 2010; Bokelmann & Lessmann, 2024; Bastani et al., 2026).

We call this failure mode **acquisition collapse**: a costly signal appears valuable under aggregate or post hoc evaluation, yet its value does not translate into reliable held-out decisions about when to acquire it (Figure 1). The distinction is fundamental because acquisition decisions must be made *before* the expensive observation is available. A signal can therefore be useful without being actionable.

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

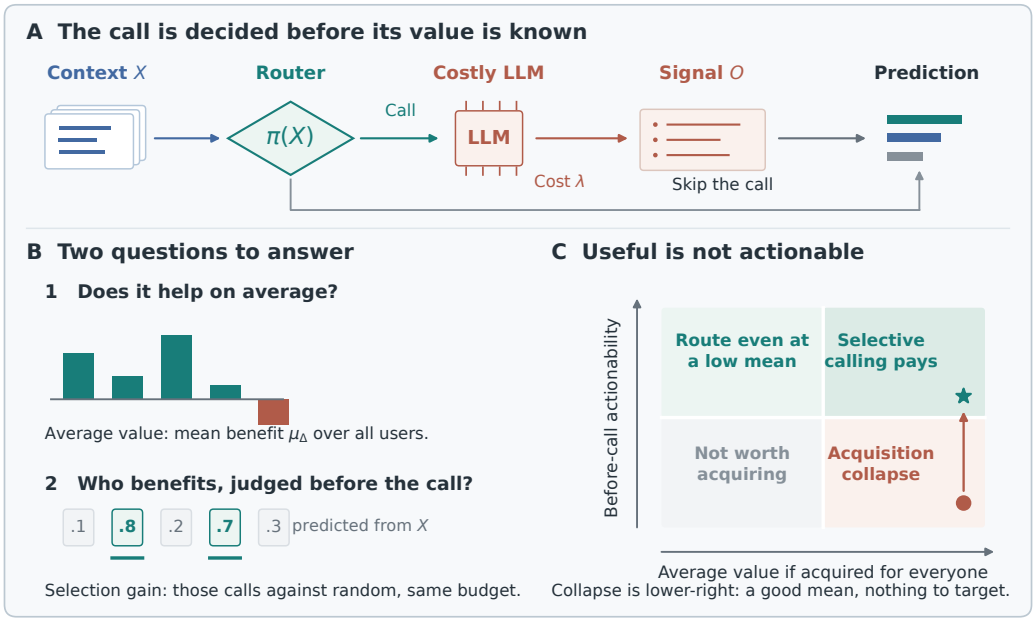


Figure 1: **Acquisition collapse: positive average value without predictable selective value. A:** From cheap X, the router must estimate the incremental value of costly signal O before paying for it. **B:** Average value asks whether acquisition helps overall; before-call predictability asks where it helps. Selection gain G_b measures selective value against random acquisition at the same budget. **C:** These two axes define four regimes (schematic). Acquisition collapse is lower right: positive average value without before-call predictability.

To understand and diagnose acquisition collapse, this paper addresses two questions. **(Q1) How can we tell whether a costly signal is selectively actionable?** We introduce **RAISE** (Reward-SNR Actionability in Signal Evaluation), a pre-routing diagnostic framework that separates overall signal value from value that can be predicted from before-call information. RAISE combines reward precision, predictable heterogeneity, independently credited policy value, and acquisition cost into a common audit. Its purpose is deliberately upstream of router design: before optimizing how to route an expensive model, we ask whether the available evidence supports routing at all. **(Q2) When does selective acquisition actually become deployable?** We instantiate RAISE with **Structured Hypothesis Embeddings** (SHE), a frozen-LLM intent signal for recommendation requiring one LLM call per user. We study it through controlled, retrospective, fresh-cohort, and prospective evaluations, with prospective audit decisions frozen before independent outcomes are revealed. Across these settings, **predictable incremental benefit, not average lift alone, is what determines whether selective LLM calling becomes deployable**. Apparent oracle or subgroup gains can vanish under independent evaluation, while selective value emerges when before-call context contains sufficient information about where acquisition helps.

This perspective connects LLM routing to the broader problem of **active information acquisition** (Ma et al., 2019; Li & Oliva, 2021). Tool calls, stronger-model consultation, robotic sensing, and costly clinical measurements all require a system to pay for information before knowing its value. RAISE provides a common way to ask whether that value is predictable enough to justify selective acquisition.

Our contributions are threefold.

1. **Acquisition collapse.** We identify and formalize a failure mode that separates the usefulness of a costly signal from its actionability under a limited acquisition budget.
2. **RAISE.** We introduce a general pre-routing framework for testing whether before-call information supports reliable selective acquisition before investing in a routing policy.

3. **Evidence across controlled and deployment-style settings.** Using SHE as a concrete costly LLM signal, we show across controlled, retrospective, fresh-cohort, and prospective studies that selective value tracks predictable incremental benefit rather than average lift alone.

2 ACQUISITION COLLAPSE AND THE RAISE AUDIT

RAISE turns the introduction's distinction into an audit by separating two properties of a costly signal: its average value when acquired and whether its incremental value is predictable before acquisition.

For each example, X is cheap before-call context, O an optional costly observation, and $R(1), R(0)$ paired offline rewards from the pipeline with O (augmented) and without it (base). Let $\Delta = R(1) - R(0)$: $\mu_\Delta = \mathbb{E}[\Delta]$ is average value and $\tau(x) = \mathbb{E}[\Delta \mid X = x]$ is predictable incremental benefit. A policy $\pi(X) \in \{0, 1\}$ sees only X; its evaluation rewards are held out from fitting and selection. These offline contrasts require separate deployment validation.

Definition (acquisition collapse). Fix a cohort, reward construction, before-call features, budget, sample size, and policy family. We say that *acquisition collapse* occurs in this setting when a positive average-use effect, or a post hoc oracle gain from choosing examples after seeing their rewards, is not matched by an established positive held-out *selection gain* G_b (Eq. 2) among the evaluated policies. A confidence interval that crosses zero leaves selection unresolved, whereas an upper bound below zero supports worse-than-random selection.

The same policy is the acquisition gate of an agent: $\pi(X)$ decides whether to acquire O before a fixed downstream rule acts; in our experiments, it decides whether to generate SHE before ranking.

2.1 MEASURING SELECTIVE VALUE

When only a fraction b of examples can receive the signal, random acquisition spends that budget ignoring X, while a budgeted policy $\pi_b(X)$ targets larger predicted benefit. With a constant per-call cost λ expressed in reward units, net utility relative to skipping is

$$U(\pi) = \mathbb{E}[\pi(X)\Delta] - \lambda\mathbb{E}[\pi(X)]. \tag{1}$$

At budget $\mathbb{E}[\pi_b] = b$, random acquisition has utility $b\mu_\Delta - b\lambda$. We define *selection gain* as the policy's additional utility over random acquisition at the same budget:

$$\begin{aligned} G_b(\pi_b) &= U(\pi_b) - U(\text{random}_b) = \mathbb{E}[(\pi_b(X) - b)\Delta] \\ &= \text{Cov}(\pi_b(X), \tau(X)). \end{aligned} \tag{2}$$

Thus $G_b > 0$ means that the policy preferentially acquires examples whose conditional benefit exceeds the population average; under constant per-call cost, cost cancels from G_b.

Eq. 2 is the population average prescriptive effect (PAPE) at treatment rate b (Imai & Li, 2023); G_b/b is the selected-group effect relative to the population mean, the targeting operator characteristic of the rank-weighted average treatment effects (RATE) framework (Yadlowsky et al., 2025). We use G_b as the endpoint.

Selective value is invariant to a uniform benefit shift. Adding the same constant c to every incremental reward leaves a fixed rate-b policy's gain unchanged: $\mathbb{E}[(\pi_b - b)(\Delta + c)] = G_b$. Balanced $X \in \{-1, +1\}$ with $\tau(X) = aX$, $a > 0$, has zero mean but $G_{.5} = a/2$ when acquiring at $X = +1$; constant $\tau(X) = \mu > 0$ gives $G_b = 0$ for every rate-b policy. Average value and actionability therefore vary independently.

2.2 THE RAISE AUDIT

RAISE asks four questions in order: What is compared? Is the reward signal measurable? Is benefit predictable before acquisition? Does held-out policy value justify its cost? An audit sample that acquires O for every example, observing both rewards, answers them while keeping apart the measurements, the frozen rule that reads them, and the router.

Protocol 1 RAISE: from pre-routing diagnostics to deployment evidence

Inputs: paired rewards $R(0), R(1)$; before-call covariates X; budget b; benefit target ϵ; cost model; policies.

1. **Fix the comparison.** Record the cohort, sampling unit, reward pipeline, feature allowlist, splits, tuning rules, and planned inference, and keep evaluation rewards outside fitting and selection.
2. **Characterize the channel.** Measure signed average value, reward signal-to-noise ratio (reward-SNR), precision, and feature-predictable heterogeneity against appropriate nulls and positive controls; freeze prospective decisions before outcomes.
3. **Credit selection independently.** Fit policies within training data, evaluate G_b on held-out rewards, state the scope of intervals, and separate perturbation references from tests with a justified null.
4. **Compare acquisition choices.** Report evidence for $G_b > 0$, attainment of ϵ, and net utility against random acquisition, acquire-all, and skip, preserving unresolved comparisons.

Output: diagnostics, intervals, and supported decisions, tied to the cohort, reward pipeline, budget, and cost.

Step 2 first asks whether average value is measurable. Reward-SNR $\rho = |\mu_\Delta|/\sigma_\Delta$ describes average-effect precision. We report the sign of μ_Δ and $N_{\min} \approx (2.8/\rho)^2$, the sample size at which a two-sided $\alpha = .05$ test has 80% planned power for the mean. $N_{\min}$ calibrates mean-effect precision (Appendix F). To ask whether that value can be targeted, we use heterogeneous treatment effect (HTE) diagnostics: out-of-fold (OOF) score–reward correlation, prediction R^2, and group average treatment effects (GATES)-style contrasts (Chernozhukov et al., 2018); matched positive controls with injected signal establish sensitivity. Permutation tests are valid when they recompute the statistic under a justified transformation (Ojala & Garriga, 2010; Hemerik & Goeman, 2018); refitting on disrupted rewards and crediting on real ones is only a stability reference. These diagnostics indicate whether targeting structure exists; held-out G_b measures whether a policy recovers it.

Step 4 turns evidence into decisions, which require both a practical target and a cost model. Against target $\epsilon > 0$, interval $[L, H]$ is *confirmed-useful* if $L > \epsilon$, *below-target* if $H \leq \epsilon$, and *unresolved* otherwise. Crossing ϵ after clearing zero means positive gain but unresolved target attainment. By Eq. 1, net value at budget b is $b(\mu_\Delta - \lambda) + G_b$, so a positive gain may still be uneconomic; variable costs require comparing expected expense. With uniform cost and no overhead, selection beats acquire-all and skip only if its net value exceeds $\max\{\mu_\Delta - \lambda, 0\}$ (Appendix G).

3 Experimental Instantiation and Study Design

RAISE needs paired outcomes with and without the costly signal. Recommendation provides them because base and LLM-augmented pipelines can score the same users. We use Structured Hypothesis Embeddings (SHE) as the costly observation O. Given a user history, a frozen LLM produces $K = 3$ intent hypotheses h_k, confidence scores γ_k, and indices that point to supporting history events. With frozen hypothesis embeddings $\boldsymbol{e}_k = \phi(h_k)$ and candidate embedding $\boldsymbol{c}$, we summarize the best confidence-weighted match as

$$f_B^{\max}(\boldsymbol{c}) = \max_{k=1,\ldots,K} \gamma_k \cos(\boldsymbol{c}, \boldsymbol{e}_k). \tag{3}$$

This score is one coordinate of a four-feature SHE block. A late-fusion ranker concatenates that block with the base recommender score; Appendix H details the full block, encoders, and generator dependence. We then test selective value in four stages, each removing a source of optimistic bias (Table 1): a synthetic controlled study (Section 4) followed by three real-data studies (Section 5).

Table 1: **Four stages, four questions.** Each stage strengthens a distinct separation.

Stage	Key separation	Question
Controlled	Aligned/permuted features; fixed rewards	Does predictability move G_b without changing the mean?
Original	Cross-fitted rewards; independent policy credit	Do oracle or subgroup gains transfer?
Fresh MIND	Frozen upstream models; OOF routers	Does held-out selection emerge?
Prospective	Disjoint audit/train/evaluation; frozen decisions	Do frozen predictions transfer?

We begin retrospectively. On the original Microsoft News Dataset (MIND) (Wu et al., 2020) ($N = 1263$) and the Amazon Reviews 2023 All_Beauty cohort (Amazon-Beauty; $N = 650$) (Hou et al., 2024), cross-fitted rewards replace *self-credit*, in which the same rewards both choose and score a selection: each evaluation unit is scored by reward models fit without its fold. Fresh MIND then adds $N = 4954$ unseen users while keeping prior-cohort-fitted latent semantic analysis (LSA) representations, gated recurrent unit (GRU) and self-attentive sequential recommendation (SASRec) backbones (Hidasi et al., 2016; Kang & McAuley, 2018), and ranking heads frozen; routers are trained and credited out of fold. The stages use different populations and reward pipelines (Appendices A, K, L, and H.2).

Those studies remain retrospective. The final gap is temporal: can the audit commit before outcomes are known? The *prospective pilot* tests this on new MIND and Amazon Reviews 2023 Beauty and Personal Care users (Hou et al., 2024). It is prospective because audit decisions and policy selections are frozen before evaluation rewards are scored, and a pilot because its four conditions (GRU and SASRec on each domain) can test transfer but are too few to estimate the framework's error rates. MIND GRU is the sole pre-specified primary endpoint; MIND SASRec and both Amazon conditions are secondary. Each domain assigns 800, 1,600, and 1,600 users to audit, router training, and evaluation. Backbones share users and generated signals; historical reward models remain frozen. We evaluate the fitted policy in all four conditions regardless of the audit decision. The router sees ten before-call features: candidate and history counts, four history–candidate similarity summaries, two normalized backbone-score summaries, and sparse-history and multi-intent indicators; SHE outputs enter only the ranker (Appendices A, E, and E.2, including GPT-5.5 provenance limits).

Uncertainty follows each stage's sampling unit. In the prospective pilot, the endpoint is $G_{.30}$ over budget-matched random acquisition at a 30% call budget. Ridge policies fit on router users and select 480 of the 1,600 evaluation users before reward scoring. Ten thousand shared-user bootstrap draws give fixed-policy intervals conditional on fitted models and selections; we compare them with the pre-specified target $\epsilon = .005$ in normalized discounted cumulative gain at rank 10 (NDCG@10) (Järvelin & Kekäläinen, 2002). Fresh MIND uses five-fold OOF policies with refit sensitivities, whereas controlled intervals span draw seeds.

4 Controlled Evidence: Predictability Creates Selective Value

Before turning to real cohorts, a synthetic Gaussian study tests the prediction of Section 2: selection gain follows predictable heterogeneity, not average benefit. Each user has eight standard-Gaussian before-call features; true benefit is a common mean plus a term along one fixed feature direction, and independent Gaussian noise produces training and evaluation rewards. Three Gaussian cases vary these two ingredients. In the constant-positive case, everyone has the same positive benefit, so acquisition helps on average but offers no better target. The positive-mean heterogeneous case adds predictable user-level differences. The zero-mean heterogeneous case removes only the common mean shift, preserving those differences around zero (Appendix B).

Within each case, the aligned arm gives routers the true features, while the permuted arm shuffles feature rows across users; both arms share rewards, users, and folds, so only feature–benefit alignment changes. If predictability matters, permutation should erase G_b, while constant-positive should show none. Policies train on one noise realization and are credited on an independent one (*noisy credit*); *latent credit* instead scores them against true benefit. Settings were fixed on three development draws; the primary comparison is aligned-minus-permuted gain across 20 separate paired validation draws.

That is what we observe. Relative to permuted features, aligned features add $+0.10556$ noisy-credit selection gain in both heterogeneous cases, with 95% interval $[+0.10179, +0.10933]$; constant-positive has average value but gives -0.00149 $[-0.00595, +0.00298]$ selection gain (Figure 2). Identical gains confirm invariance to the mean shift.

The zero-mean case exposes the cost of requiring average value and heterogeneity together. A conjunctive screen approves only when both a positive-mean test and an HTE test pass. The zero-mean aligned arm passes the HTE screen in 20/20 draws but the positive-mean screen in only 1/20, so the conjunctive screen withholds approval in 19/20 despite recoverable selective value. This miss is structural: larger samples sharpen inference around zero while the aligned heterogeneity persists.

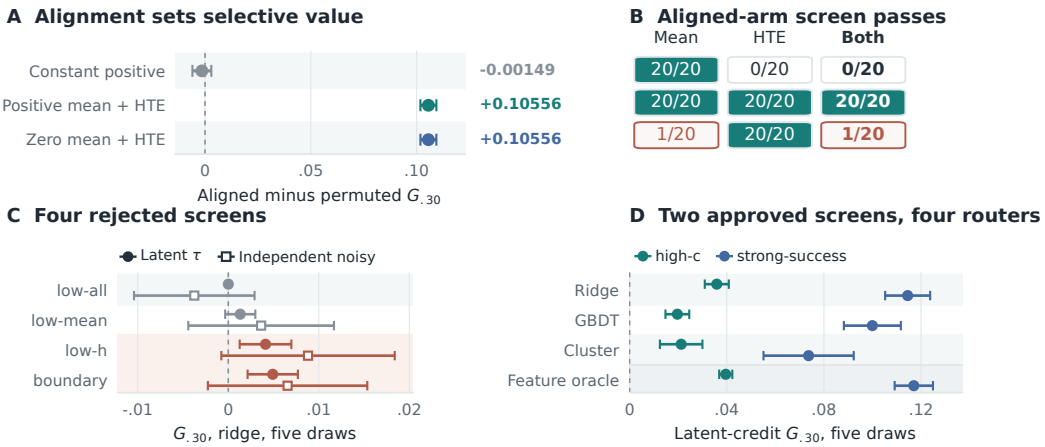


Figure 2: **Predictable incremental benefit drives selective value.** Values are in synthetic reward units; each panel has its own x scale. **A:** Aligned and permuted features share rewards; only alignment changes. Points show independent-noisy-credit G_b with paired Student-t 95% intervals over 20 draws; the heterogeneous cases differ by a constant mean shift, so their gains coincide. **B:** Aligned-arm pass counts: requiring both mean and HTE screens withholds approval from the zero-mean case in 19/20 draws. **C:** Frozen six-case scan (five draws); two shaded false negatives are positive under latent but unresolved under noisy credit. **D:** Every router from the two approved cases; the feature oracle is an upper-bound reference.

A six-case scan checks the pattern beyond this construction: six further Gaussian cases, five draws each, screened by a conjunctive rule frozen before outcomes. Its two approved cases retain independent-noisy-credit gains of 0.03624 $[0.02548, 0.04699]$ and 0.11705 $[0.10160, 0.13250]$; all three learned routers (ridge, gradient-boosted trees, and a cluster router) recover value under both latent and noisy credit (Appendix B).

5 Real-Data Evidence: From Collapse to Prospective Transfer

We now test whether the controlled pattern survives the three real-data stages of Table 1. Figure 3 tracks the first two and their cost implications.

5.1 Original cohorts: apparent gain fails independent credit

We start with self-credit on original MIND. A self-credited oracle, which selects the examples with the highest realized rewards, still gains $+0.0518$ after reward permutation because one reward realization both ranks and credits examples (Figure 3A). A cluster mined for high reward gains $+0.0097$ when the same rewards credit it but falls to -0.0002 under independent credit; a selector trained on noise and credited on real rewards yields -0.0041. Positive OOF selection remains unresolved across the evaluated policy family: apparent value fails independent credit, the signature of acquisition collapse.

Could a different router exploit heterogeneity missed by these policies? On MIND, OOF score–reward correlation is -0.0181, $R^2 = -0.0102$, and the grouped top-minus-bottom contrast is $+0.00254$. HTE-SNR, the standard deviation of OOF predictions divided by residual standard deviation, is 0.0829. Every statistic, like its Amazon counterpart, falls inside its permutation-null band (Figure 3B); a sensitivity analysis with reward models fit on disjoint users leaves R^2 and the grouped contrast unresolved. Positive controls that inject known effects into eight k-means clusters (KMeans-8) are first recovered at cluster-SNR, between- over within-cluster reward standard deviation, of .20 on MIND and .35 on Amazon, above the real .0799 and .0628, so sensitivity is limited at these sizes (Appendix C). Two explanations remain: no predictable signal or too little data.

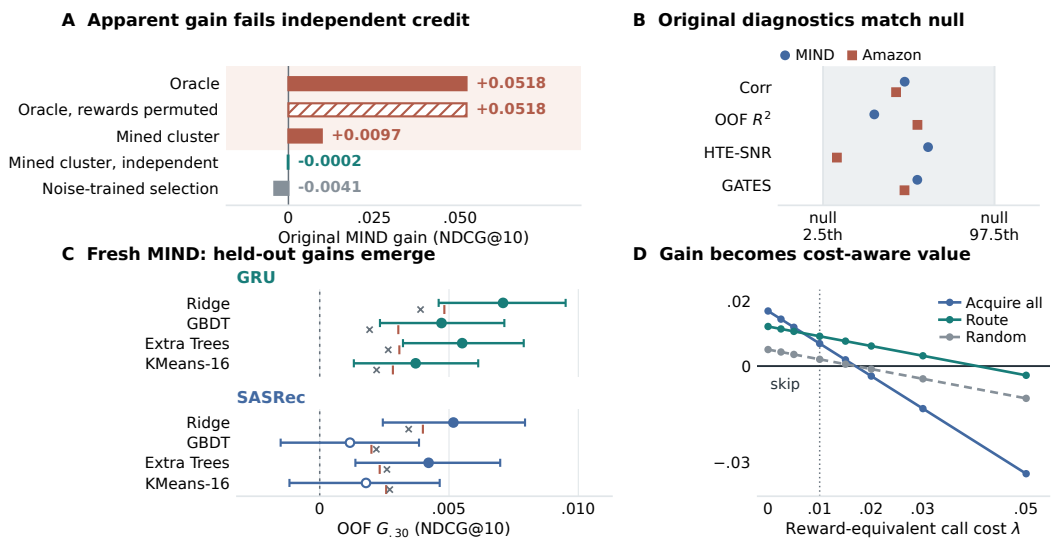


Figure 3: **Independent credit separates apparent gain from held-out selective value.** Cohorts and reward pipelines differ; all intervals are 95%. **A:** On original MIND ($N = 1263$), a self-credited oracle survives reward permutation, while a mined cluster loses its gain under independent credit; shading marks reused sorting rewards. **B:** All eight original MIND/Amazon diagnostics lie inside their own permutation-null bands. **C:** On fresh MIND ($N = 4954, b = .30$), six of eight OOF policies have positive policy-by-policy (marginal) intervals; ticks and crosses are uncalibrated perturbation references. **D:** GRU-ridge routing beats acquire-all only beyond the dotted cost; skip wins at high cost.

5.2 Fresh MIND: held-out selection emerges

Fresh MIND probes the data-limitation explanation, although its cohort and reward pipeline also change. With $N = 4954$ unseen users, frozen upstream models, and OOF router credit, selective value becomes detectable. Six of eight policies, all reported in this exploratory stage, have fixed-policy intervals entirely above zero; these marginal intervals are not adjusted across policies (Figure 3C). Ridge gains are 0.00709 $[0.00460, 0.00951]$ for GRU and 0.00517 $[0.00244, 0.00794]$ for SASRec. Average lift stays positive across historical cohorts of increasing size, while positive policy intervals emerge only at larger sizes (Figure 7). Ten acquisition-split refits and five reward-pipeline conditions keep 59/60 and 28/30 nominal intervals above zero as correlated stability summaries (Appendix D). Average-benefit detection and selective-policy learning therefore emerge on different sample-size scales. The 1,000-draw permuted-reward and matched-noise refits in Figure 3C lack an exchangeability justification for calibrated null tests (Appendix D.1).

With 1,600 evaluation users fixed and router-training size varied from 100 to 3,354, both maximum-size ridge intervals exclude zero. The nonmonotonic curves and contrasts between maximum and 500-user training sizes, which cross zero, place onset at dataset- and model-specific scales (Figure 6; Appendix D.2).

When is selection worth paying for? At approximately 30% calls, the selected increment $V_b = \mathbb{E}[\pi_b \Delta] = b\mu_\Delta + G_b$ retains 71.86% $[58.30\%, 89.22\%]$ of the always-acquire increment with GRU and 68.51% $[48.59\%, 99.12\%]$ with SASRec, under joint fixed-policy bootstrap. Acquire-all is preferable while calls are nearly free, routing becomes preferable at nontrivial cost, and skipping dominates when acquisition is expensive (Figure 3D). Costs are in reward-equivalent units (Appendix G).

5.3 Prospective pilot: frozen decisions, open rule comparison

Fresh MIND establishes held-out value, but not whether an audit would predict it before the final outcomes. We froze six rules or references. Mean-only approves a significantly positive audit mean; HTE-only approves when a permutation-tested HTE statistic passes. RAISE-v1 conservatively

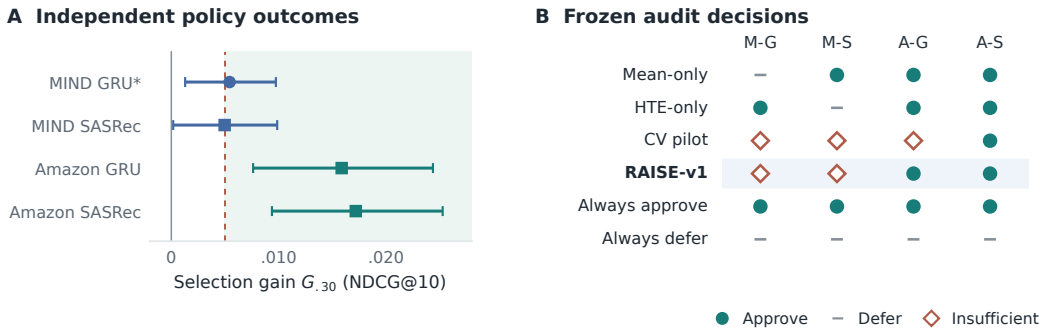


Figure 4: **Frozen audit decisions and independent policy outcomes answer different questions.** **A:** Four prospective ridge endpoints at $b = .30$ with 95% fixed-policy intervals and target $\epsilon = .005$ dashed. MIND GRU is the sole primary: MIND clears zero but not the target; correlated Amazon secondaries clear it. **B:** Six rules frozen before outcomes; M/A and G/S denote dataset and backbone. Intervals condition on fitted policies and locked selections.

conjoins them, acting only when both agree and otherwise recording insufficient evidence; the broader framework treats these dimensions separately. A small cross-validated (CV) router audit ("CV pilot" in Figure 4) checks its gain interval against ϵ; always-approve and always-defer are constant references. All four data-dependent rules share one 800-user audit sample (Appendix E).

On MIND, both fixed-policy intervals exclude zero: the primary GRU endpoint gains $G_{.30} = 0.00543$ (95% interval $[0.00129, 0.00974]$) and SASRec 0.00497 $[0.00017, 0.00986]$ (Figure 4). Both cross the pre-specified target $\epsilon = .005$, leaving practical usefulness unresolved. RAISE-v1 had recorded insufficient evidence for both. The component screens disagree: mean-only approves SASRec, whereas HTE-only approves GRU; against the target, these approvals remain unclassified.

On Amazon, both correlated secondary endpoints exceed the target: GRU gains 0.01587 $[0.00762, 0.02437]$ and SASRec 0.01718 $[0.00937, 0.02528]$. RAISE-v1, mean-only, HTE-only, and always-approve had all approved both; the CV pilot approved SASRec and abstained on GRU.

The pilot establishes that decisions frozen before outcomes can be checked against later outcomes, not that any rule is superior. Simpler rules make the same successful approvals, leaving the rules tied on four outcomes with no resolved below-target condition. A broader prospective study spanning useful and below-target settings is needed to compare false approvals and quantify RAISE-v1's incremental value.

6 Related Work

LLM routing and cost-aware inference. Recent work reduces LLM inference cost through cascades, model routing, and selective invocation. FrugalGPT learns cascades across models with different costs and capabilities (Chen et al., 2024), RouteLLM learns query-level routing from preference data (Ong et al., 2025), and later work unifies the two (Dekoninck et al., 2025). RouterBench and LLMRouterBench standardize evaluation of routing systems (Hu et al., 2024; Li et al., 2026). These methods ask how to route once a routing problem has been posed. RAISE asks an upstream question: does the information available before the costly call contain enough predictable incremental value to justify selective routing at all? The distinction matters when apparent average or oracle gains do not translate into deployable selection.

Active information acquisition, deferral, and selective prediction. Active feature acquisition chooses which measurements to obtain from partially observed inputs (Ma et al., 2019; Li & Oliva, 2021), while learning-to-defer and selective prediction decide when to rely on an alternative predictor or expert (Madras et al., 2018; Mozannar & Sontag, 2020; Geifman & El-Yaniv, 2017); conformal routing constrains violations against an expensive-model reference (Uddin & Bauer, 2026). RAISE is complementary: instead of proposing another acquisition policy, it tests whether before-call

information supports selective acquisition under a stated budget and cost model. It can therefore precede many routing, deferral, or acquisition algorithms.

Heterogeneous effects and policy evaluation. Our distinction between average and selective value connects to heterogeneous treatment-effect estimation and policy learning. Effect estimation and decision quality are not the same objective (Fernández-Loría & Provost, 2022). Empirical welfare maximization and policy learning target decisions directly (Kitagawa & Tetenov, 2018; Athey & Wager, 2021; Bastani et al., 2025). PAPE and RATE quantify improvement from targeting over non-targeted allocation (Imai & Li, 2023; Yadlowsky et al., 2025), while GATES-style analyses inspect predictable effect variation (Chernozhukov et al., 2018). RAISE adopts these ideas for costly signal acquisition, using held-out selection value to ask whether before-call context identifies where an optional observation is most useful. Our contribution is not a new HTE estimator, but a deployment-oriented audit that combines mean value, predictable heterogeneity, independent policy credit, and acquisition cost in one pre-routing evaluation.

Diagnosing hidden failures in LLM systems. Aggregate performance can also conceal the dependence that deployment needs. RAGEN-2 diagnoses reasoning collapse in agentic RL by showing that aggregate indicators can miss the relationship between inputs and useful reasoning behavior (Wang et al., 2026). RAISE studies a different dependence: whether information available before acquisition predicts the incremental value of a costly signal. The shared lesson: aggregate success is insufficient when deployment depends on a finer-grained relationship. Acquisition collapse names this failure for costly acquisition and gives a way to test for it before routing is deployed.

7 Conclusion and Limitations

Conclusion. Costly LLM signals create a deployment problem that average performance alone cannot answer: even when an LLM helps, a system still needs to know when the call is worth paying for. We identify **acquisition collapse**, in which a signal appears valuable in aggregate or post hoc but provides too little before-call information for reliable selective acquisition. We introduce **RAISE** (Reward-SNR Actionability in Signal Evaluation) to diagnose this gap by separating overall signal value from value that is predictable and actionable before acquisition. Using **Structured Hypothesis Embeddings** (SHE) as a frozen-LLM signal for recommendation, we evaluate RAISE through controlled, retrospective, fresh-cohort, and prospective studies. Across these settings, **predictable incremental benefit, not average lift alone, is what makes selective acquisition actionable**: apparent oracle or subgroup value can disappear under independent evaluation, while selective value emerges when before-call context contains useful information about where acquisition helps. More broadly, our results recast costly LLM invocation as a problem of actionable information acquisition: before building a router around an expensive model, tool, sensor, or measurement, first test whether its value is predictable at decision time.

Limitations. Our current evidence focuses on single-step recommendation and one LLM-signal design, motivating broader validation across models, tasks, and domains. The prospective pilot validates frozen-before-outcome auditing and positive held-out selection, but does not yet establish when RAISE-v1 adds value beyond simpler screens. Production studies should incorporate measured monetary cost, latency, routing overhead, and acquisition failures rather than reward-equivalent costs alone. Extending RAISE to agents, robotics, and other sequential settings will additionally require history-dependent rewards, trajectory-level evaluation, and accounting for how one acquisition changes later decisions and future calls (Appendix O).

AI Use Statement

The experimental method uses a frozen LLM to generate structured intent hypotheses from histories, as described in Section 3 and Appendix H. AI tools also assisted with literature search.

Reproducibility Statement

The real-data studies use public recommendation datasets. We document reward definitions, data roles, frozen rules, prompts, and uncertainty procedures in the main text and appendices. Preserved

experiment packages contain source snapshots, manifests, cached outputs, and offline reconstruction scripts; the prospective endpoint and bootstrap calculations have been independently reconstructed. This verification covers cached numerical evidence, with upstream reproduction and historical provenance limits described in Appendix N. Code will be released upon acceptance.

REFERENCES

Linus Aronsson, Arman Rahbar, and Morteza Haghir Chehreghani. A survey on active feature acquisition strategies. *arXiv preprint arXiv:2502.11067*, 2025.

Susan Athey and Stefan Wager. Policy learning with observational data. *Econometrica*, 89(1): 133–161, 2021. doi: 10.3982/ECTA15732.

Hamsa Bastani, Osbert Bastani, and Bryce McLaughlin. Beating the winner's curse via inference-aware policy optimization. *arXiv preprint arXiv:2510.18161*, 2025. URL https://arxiv.org/abs/2510.18161.

Hamsa Bastani, Osbert Bastani, and Bryce McLaughlin. Winner's curse drives false promises in data-driven decisions: A case study in refugee matching. *arXiv preprint arXiv:2602.08892*, 2026. URL https://arxiv.org/abs/2602.08892.

Björn Bokelmann and Stefan Lessmann. Improving uplift model evaluation on randomized controlled trial data. *European Journal of Operational Research*, 313(2):691–707, 2024. doi: 10.1016/j.ejor.2023.09.018.

Gavin C. Cawley and Nicola L. C. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11(70):2079–2107, 2010.

Lingjiao Chen, Matei Zaharia, and James Zou. FrugalGPT: How to use large language models while reducing cost and improving performance. *Transactions on Machine Learning Research*, 2024.

Victor Chernozhukov, Mert Demirer, Esther Duflo, and Iván Fernández-Val. Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in india. Working Paper 24678, National Bureau of Economic Research, 2018.

Yu-Neng Chuang, Prathusha Kameswara Sarma, Parikshit Gopalan, John Boccio, Sara Bolouki, Xia Hu, and Helen Zhou. Learning to route LLMs with confidence tokens. In *International Conference on Machine Learning (ICML)*, pp. 10859–10878, 2025. URL https://proceedings.mlr.press/v267/chuang25b.html.

Jasper Dekoninck, Maximilian Baader, and Martin Vechev. A unified approach to routing and cascading for LLMs. In *International Conference on Machine Learning (ICML)*, pp. 12987–13010, 2025. URL https://proceedings.mlr.press/v267/dekoninck25a.html.

Dujian Ding, Ankur Mallick, Shaokun Zhang, Chi Wang, Daniel Madrigal, Mirian Del Carmen Hipolito Garcia, Menglin Xia, Laks V. S. Lakshmanan, Qingyun Wu, and Victor Rühle. BEST-Route: Adaptive LLM routing with test-time optimal compute. In *International Conference on Machine Learning (ICML)*, pp. 13870–13884, 2025. URL https://proceedings.mlr.press/v267/ding25d.html.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *International Conference on Machine Learning (ICML)*, 2011. URL https://icml.cc/2011/papers/554_icmlpaper.pdf.

Tao Feng, Fangxu Yu, Haozhen Zhang, Zhongjie Dai, Liangqi Yuan, Zijie Lei, Weizhi Zhang, Kunlun Zhu, Haodong Yue, Keyang Xuan, Ge Liu, and Jiaxuan You. LLMRouter: Unified infrastructure for developing, evaluating, and deploying LLM routers. *arXiv preprint arXiv:2608.06867*, 2026.

Carlos Fernández-Loría and Foster Provost. Causal decision making and causal effect estimation are not the same...and why it matters. *INFORMS Journal on Data Science*, 1(1):4–16, 2022. doi: 10.1287/ijds.2021.0006.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Jesse Hemerik and Jelle Goeman. Exact testing with random permutations. *TEST*, 27:811–825, 2018. doi: 10.1007/s11749-017-0571-1.

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

Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. Bridging language and items for retrieval and recommendation. *arXiv preprint arXiv:2403.03952*, 2024.

Qitian Jason Hu, Jacob Bieker, Xiuyu Li, Nan Jiang, Benjamin Keigwin, Gaurav Ranganath, Kurt Keutzer, and Shriyash Kaustubh Upadhyay. RouterBench: A benchmark for multi-LLM routing system. *arXiv preprint arXiv:2403.12031*, 2024.

Kosuke Imai and Michael Lingzhi Li. Experimental evaluation of individualized treatment rules. *Journal of the American Statistical Association*, 118(541):242–256, 2023. doi: 10.1080/01621459.2021.1923511.

Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002. doi: 10.1145/582415.582418.

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

Toru Kitagawa and Aleksey Tetenov. Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616, 2018. doi: 10.3982/ECTA13288.

Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019. doi: 10.1073/pnas.1804597116.

Hao Li, Yiqun Zhang, Zhaoyan Guo, Chenxu Wang, Shengji Tang, Qiaosheng Zhang, Yang Chen, Biqing Qi, Peng Ye, Lei Bai, Zhen Wang, and Shuyue Hu. LLMRouterBench: A massive benchmark and unified framework for LLM routing. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 37733–37754, 2026.

Yang Li and Junier Oliva. Active feature acquisition with generative surrogate models. In *International Conference on Machine Learning (ICML)*, pp. 6450–6459, 2021. URL https://proceedings.mlr.press/v139/li21p.html.

Chao Ma, Sebastian Tschiatschek, Konstantina Palla, Jose Miguel Hernandez-Lobato, Sebastian Nowozin, and Cheng Zhang. EDDI: Efficient dynamic discovery of high-value information with partial VAE. In *International Conference on Machine Learning (ICML)*, pp. 4234–4243, 2019. URL https://proceedings.mlr.press/v97/ma19c.html.

David Madras, Toniann Pitassi, and Richard Zemel. Predict responsibly: Improving fairness and accuracy by learning to defer. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *International Conference on Machine Learning (ICML)*, pp. 7076–7087, 2020. URL https://proceedings.mlr.press/v119/mozannar20b.html.

Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319, 2021. doi: 10.1093/biomet/asaa076.

Markus Ojala and Gemma C. Garriga. Permutation tests for studying classifier performance. *Journal of Machine Learning Research*, 11(62):1833–1863, 2010.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M. Waleed Kadous, and Ion Stoica. RouteLLM: Learning to route LLMs with preference data. In *International Conference on Learning Representations (ICLR)*, 2025.

Xubin Ren, Wei Wei, Lianghao Xia, et al. Representation learning with large language models for recommendation. In *The Web Conference (WWW)*, 2024.

Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *International Conference on Machine Learning (ICML)*, pp. 814–823, 2015. URL https://proceedings.mlr.press/v37/swaminathan15.html.

Iqtedar Uddin and André Bauer. Conformal LLM routing with distribution-free safety guarantees. In *Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pp. 791–799, 2026. doi: 10.18653/v1/2026.acl-srw.70.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018. doi: 10.1080/01621459.2017.1319839.

Zihan Wang, Chi Gui, Xing Jin, Qineng Wang, Licheng Liu, Kangrui Wang, Shiqi Chen, Linjie Li, Zhengyuan Yang, Pingyue Zhang, Yiping Lu, Jiajun Wu, Li Fei-Fei, Lijuan Wang, Yejin Choi, and Manling Li. RAGEN-2: Reasoning collapse in agentic RL. *arXiv preprint arXiv:2604.06268*, 2026.

Wei Wei, Xubin Ren, Jiabin Tang, Qinyong Wang, Lixin Su, Suqi Cheng, Junfeng Wang, Dawei Yin, and Chao Huang. LLMRec: Large language models with graph augmentation for recommendation. In *ACM International Conference on Web Search and Data Mining (WSDM)*, 2024.

Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, and Ming Zhou. MIND: A large-scale dataset for news recommendation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 3597–3606, 2020. doi: 10.18653/v1/2020.acl-main.331.

Yunjia Xi, Weiwen Liu, Jianghao Lin, et al. Towards open-world recommendation with knowledge augmentation from large language models. In *ACM Conference on Recommender Systems (RecSys)*, 2024.

Steve Yadlowsky, Scott Fleming, Nigam Shah, Emma Brunskill, and Stefan Wager. Evaluating treatment prioritization rules via rank-weighted average treatment effects. *Journal of the American Statistical Association*, 120(549):38–51, 2025. doi: 10.1080/01621459.2024.2393466.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Ying Yuan et al. HypoGeneAgent: A hypothesis language agent for gene-set cluster resolution selection using Perturb-seq datasets. *arXiv preprint arXiv:2509.09740*, 2025.

A Detailed Cohort Boundaries

Table 2 maps each study to its evaluation population and reward pipeline. The original signal analyses and the follow-up acquisition studies use different representations, sample sizes, and outcomes. We identify the cohort and backbone with each SNR estimate so that the comparisons retain their intended meaning.

Table 2: Analysis-specific cohorts. The controlled screen and its router replication share the same experiment. Fresh MIND changes more than sample size relative to the original MIND study.

Analysis	N	Reward and evaluation scope
Original MIND acquisition	1,263	ada-002 content/late-fusion study; cross-fitted ranking rewards; corrected heterogeneity analysis.
Original Amazon-Beauty	650	Amazon Reviews 2023 All_Beauty; local LSA representation; cross-fitted base and augmented ranking rewards.
Original REES46	498	Thin-content session-purchase study; separate binary reward/probe setting, not fresh-MIND NDCG routing.
Original MIND backbone study	1,411	LSA-based ordered/unordered backbones; average-use contrast, not the fresh-MIND test cohort.
Controlled scan	4,000	Six cases, five draws each; noisy training reward and latent-effect policy credit.
Fresh MIND	4,954	One impression per user; prior-cohort-fitted, frozen LSA, GRU/SASRec, and ranking heads; realized NDCG@10 reward.
Fixed-test follow-up	1,600	One cached test set; disjoint router training ladder with 100–3,354 users.
Frozen-head MIND sensitivity	715	Heterogeneity audit after a separate 476-user reward-head fit and identity-ambiguity quarantine.
Prospective MIND	4,000	Disjoint 800/1,600/1,600 audit, router, and evaluation roles; GRU is the sole primary endpoint.
Prospective Amazon	4,000	New Beauty and Personal Care users; same role sizes; both backbone outcomes are secondary.

The prior MIND generation artifact contains 1,557 rows, of which 1,549 are valid. Analysis-specific ranking eligibility produces the smaller legacy cohorts above. Fresh MIND selects 5,000 distinct new users, with 4,954 valid generated rows. The delivered selection and overlap manifests record exclusions of both the prior analysis cohort and a previously inspected 6,000-row pool. The follow-up provides source snapshots, split hashes, and chronology records; some original generation failures and historical embedding artifacts remain unavailable. Each analysis therefore retains its own eligibility boundary.

For the original analyses, impression/window resampling may not account for repeated users or overlapping histories. Fresh MIND has one impression per user and uses user-level resampling. This addresses repeated-user dependence within the cohort but does not remove dependence induced by overlapping OOF training sets or shared news environments.

B Controlled Screen and Outcome Details

Controlled-family construction. Let τ_{lat} be the simulated per-example effect, distinguished from the conditional mean $\tau(x) = \mathbb{E}[\tau_{\mathrm{lat}} \mid X = x]$. The scale family separates a predictable component $s(X)$, an unobserved component u, and observation noise ε:

$$\tau_{\mathrm{lat}} = \mu + \sqrt{h}\,\sigma_{\mathrm{het}} s(X) + \sqrt{1-h}\,\sigma_{\mathrm{het}} u, \qquad \Delta_{\mathrm{obs}} = \tau_{\mathrm{lat}} + \varepsilon. \tag{4}$$

Figures and tables use the short label τ for latent credit. For independent, centered, unit-variance s, u, ε and positive μ, setting $\sigma_{\mathrm{het}} = c\mu$ in Eq. 4 gives $\rho = \mu/\sqrt{1 + c^2\mu^2}$. Hence, when $c\rho < 1$,

$$\mu = \frac{\rho}{\sqrt{1 - c^2\rho^2}}, \qquad \sigma_{\mathrm{sig}} = \sqrt{h}\sigma_{\mathrm{het}} = \sqrt{h}\frac{c\rho}{\sqrt{1-(c\rho)^2}}. \tag{5}$$

Eq. 5 explains how the controlled family varies predictable signal: its dispersion is tied to its mean. Other acquisition problems can vary these quantities independently. The resulting scale is useful for constructing the scan; a learner's finite-sample behavior still depends on the feature representation, policy class, and estimation procedure.

Corrected six-case scan. The controlled scan uses $N = 4000$, seeds 10000–10004, and budget $b = 0.30$. The corrected source removes the post-call fields `b_conf` and `b_max`; the eight retained covariates are available before acquisition. The binary screen requires a positive mean above its planning reference and an HTE test passing either the R^2 or grouped-contrast branch. Case-level PASS requires this conjunction in at least three of five draws. This frozen six-case rule differs from the three-way prospective agreement rule in Appendix E.

Figure 2C,D draws these five-draw outcomes: the four rejected cases under both credit endpoints, and every router fitted in the two approved cases. All cases and matched-noise controls are retained, and these results precede the independent-noisy-credit and 20-draw mechanism follow-ups. Tables 3 and 4 list the same estimates numerically.

Table 3: All frozen screen labels and primary ridge outcomes. The screen uses noisy observed reward; policy credit uses latent τ. Positive intervals in `low-h` and `boundary` are false negatives of the reported screen and are not relabeled using cost.

Case	Screen	$\rho/\rho^\star$	HTE-SNR	$G_{0.30}$ on τ	95% CI
low-all	FAIL	0.282	0.0420	$+0.00000$	$[-0.00022, +0.00022]$
low-mean	FAIL	0.532	0.0438	$+0.00132$	$[-0.00035, +0.00299]$
low-h	FAIL	2.031	0.0479	$+0.00411$	$[+0.00126, +0.00697]$
boundary	FAIL	1.032	0.0490	$+0.00491$	$[+0.00213, +0.00769]$
high-c	PASS	2.031	0.1101	$+0.03589$	$[+0.03095, +0.04084]$
strong-success	PASS	4.034	0.3061	$+0.11447$	$[+0.10520, +0.12373]$

The mean observed rewards for `low-all`, `low-mean`, `low-h`, `boundary`, `high-c`, and `strong-success` are respectively 0.01260, 0.02368, 0.09043, 0.04585, 0.09150, and 0.19090. The corresponding mean per-draw $N_{\min}$ values are 124,630; 58,189; 1,037; 4,898; 1,036; and 250. These average nonlinear plug-in estimates across draws, so they need not equal $(2.8/\bar{\rho})^2$ calculated from the displayed mean SNR. The `high-c` case targets $\rho/\rho^\star = 2$, $h = 0.5$, $c = 2$; the `strong-success` case targets 4, 1, and 2.

Table 4: Multiple routers recover value in the two PASS settings. These are the same draws used by the screen, not an independent replication.

Case	Router	Latent-credit gain	95% CI
high-c	Ridge	$+0.03589$	$[+0.03095, +0.04084]$
high-c	GBDT	$+0.01966$	$[+0.01476, +0.02456]$
high-c	Cluster	$+0.02123$	$[+0.01248, +0.02998]$
strong-success	Ridge	$+0.11447$	$[+0.10520, +0.12373]$
strong-success	GBDT	$+0.09993$	$[+0.08820, +0.11166]$
strong-success	Cluster	$+0.07370$	$[+0.05513, +0.09226]$

Ridge recovers 90.5% of the feature-oracle value in `high-c`, with 95% CI $[83.3\%, 97.6\%]$, and 97.8% in `strong-success`, $[95.9\%, 99.7\%]$. The absolute-utility sensitivity uses $\lambda = 2\max(\overline{\Delta}_{\text{obs}}, 10^{-4})$, a data-derived within-case cost. The resulting utilities are $+0.00798$ $[-0.00745, +0.02341]$ and $+0.05674$ $[+0.03812, +0.07535]$, respectively. These absolute-utility intervals answer a different question from the budget-matched contrast. All six matched-noise-trained placebo intervals contain zero; their credit uses observed reward, not latent τ.

Independent noisy credit. The follow-up fixes the original policies and generates a new noise realization, $\Delta_{\text{new}} = \tau_{\text{lat}} + \varepsilon_{\text{new}}$, for evaluation. The latent effects and users are unchanged; the independence concerns outcome noise. Table 5 reports all six primary ridge outcomes. Across-draw intervals use a Student-t critical value with four degrees of freedom. The original `low-h` and `boundary` false negatives remain part of the latent analysis, while their noisier-credit intervals now cross zero.

Paired alignment intervention. The mechanism study freezes three cases after three development draws and evaluates 20 separate validation draws. Each draw compares an aligned arm with a feature-permuted arm using the same marginal rewards, noise, and evaluation users. Each draw contains

Table 5: Primary ridge outcomes under independent noisy credit. Original latent-credit false negatives are retained as historical classifications; new uncertainty does not retroactively change the frozen decisions.

Case	Screen	Latent gain	Noisy gain	Noisy 95% interval
low-all	FAIL	$+0.00000$	-0.00376	$[-0.01042, +0.00290]$
low-mean	FAIL	$+0.00132$	$+0.00363$	$[-0.00442, +0.01167]$
low-h	FAIL	$+0.00411$	$+0.00881$	$[-0.00077, +0.01839]$
boundary	FAIL	$+0.00491$	$+0.00655$	$[-0.00226, +0.01536]$
high-c	PASS	$+0.03589$	$+0.03624$	$[+0.02548, +0.04699]$
strong-success	PASS	$+0.11447$	$+0.11705$	$[+0.10160, +0.13250]$

$N = 4000$ units, eight independent standard Gaussian before-call features, and budget $b = .30$. Let $s(X)$ be the centered, standardized projection of X onto the direction $[1, -.8, .6, .4, 0, 0, 0, 0]$. The three cases use $(\mu, \sigma_{\text{het}}) = (.1, 0), (.1, .3), (0, .3)$ and

$$\tau = \mu + \sigma_{\text{het}} s(X), \qquad \Delta_{\text{obs}} = \tau + \varepsilon_{\text{obs}}, \qquad \Delta_{\text{new}} = \tau + \varepsilon_{\text{new}},$$

where the two noise terms are independent standard Gaussians. They are the constant-positive, positive-mean heterogeneous, and zero-mean heterogeneous cases, respectively. The permuted arm reassigns complete X rows while keeping τ, both reward realizations, users, and folds fixed. The primary contrast is the within-draw difference in independent-noisy-credit G_b; its interval summarizes 20 paired differences, not independent arm means. The constant-positive case gives -0.00149 $[-0.00595, 0.00298]$. Positive-heterogeneous and zero-mean heterogeneous cases both give 0.10556 $[0.10179, 0.10933]$: their constant mean shift leaves budget-matched selection unchanged. In the zero-mean case the mean gate passes 1/20 draws, the HTE gate 20/20, and their conjunction 1/20. In the constant case those counts are 20/20, 0/20, and 0/20. In the positive-heterogeneous case all three pass 20/20. These synthetic reward units are not NDCG points.

C Corrected Heterogeneity Audit on the Original Cohorts

The heterogeneity analysis fits an OOF ridge predictor to a kNN-smoothed reward target constructed within each training fold. GATES-style groups are formed within held-out folds. The base and augmented rewards themselves use a separate impression-level cross-fitting partition. This is layered OOF evaluation, not fully nested reward/heterogeneity cross-fitting. The follow-up below separates reward-head fitting from the heterogeneity sample. For OOF predictions $\widehat{\Delta}$, the reported heterogeneity signal-to-noise ratio is $\text{HTE-SNR} = \text{SD}(\widehat{\Delta})/[\text{SD}(\Delta - \widehat{\Delta}) + 10^{-12}]$, using population standard deviations and equal unit weights. It measures variation in predicted effects relative to residual reward variation; it is not a treatment-effect estimator or a group-based cluster-SNR. Figure 5 places each diagnostic on its own scale beside its permutation-null range, which Figure 3B rescales away.

Table 6: Corrected heterogeneity diagnostics, transcribed from the frozen analysis record. The reported HTE-SNR is not the cluster-SNR used by the KMeans positive control. The delivered source and cached null draws specify the estimators and their permutation implementation.

Cohort	Diagnostic	Estimate	Permutation-null 95% band	p
MIND	Correlation	-0.0181	$[-0.0909, +0.0620]$	0.532
MIND	R^2	-0.0102	$[-0.0151, +0.0013]$	0.781
MIND	HTE-SNR	0.0829	$[0.0481, 0.1049]$	0.259
MIND	GATES spread	$+0.00254$	$[-0.04163, +0.03869]$	0.483
Amazon-Beauty	Correlation	-0.0345	$[-0.1243, +0.0865]$	0.607
Amazon-Beauty	R^2	-0.00824	$[-0.02619, +0.00641]$	0.348
Amazon-Beauty	HTE-SNR	0.0618	$[0.0552, 0.1372]$	0.915
Amazon-Beauty	GATES spread	-0.00073	$[-0.02803, +0.02943]$	0.547

Table 6 lists every diagnostic. On MIND, score correlation is positive in three of five folds, R^2 in one of five, and the grouped top-minus-bottom contrast in two of five. Amazon-Beauty has negative R^2 in all five folds and varying grouped-contrast signs. The null results support no detected stable

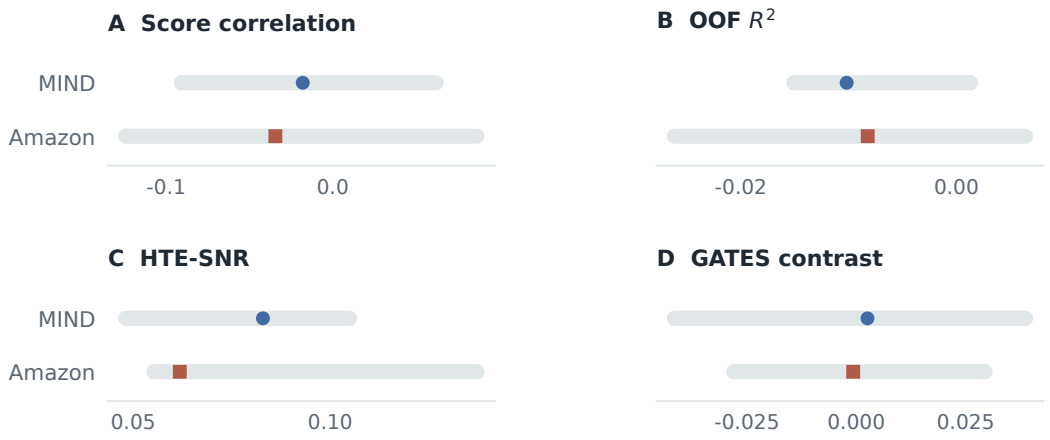


Figure 5: **All eight original-cohort heterogeneity diagnostics lie within their permutation-null bands.** Points show MIND and Amazon estimates; grey segments show the reported 95% permutation-null ranges. These are reference ranges, not confidence intervals around the points. Each metric uses its own scale.

heterogeneity for these tested estimators and samples. They do not establish absence of an effect under every feature representation or policy class.

The corrected cluster-SNR values are 0.0799 for MIND and 0.0628 for Amazon-Beauty. Here cluster-SNR is the unit-weighted between-cluster standard deviation of cluster-level effects divided by the unit-weighted within-cluster residual standard deviation; the matched controls vary this ratio by injecting cluster effects. KMeans-8 first recovers the matched injected signals at cluster-SNR 0.20 and 0.35, respectively. It measures structure available to this grouping procedure and is distinct from HTE-SNR. Global mean reward-SNR describes a different property and is reported separately in Appendix F.

Frozen-head sensitivity. The stricter follow-up fits the reward head on 476 users and audits a disjoint 715-user set. Mean reward is 0.014995, prediction $R^2 = -0.018523$ ($p = .825$), and the grouped contrast is -0.020909 ($p = .781$), using 999 permutations. This sensitivity does not recover detectable heterogeneity. It has no embedded positive-control sweep; the matched recovery controls above belong to the original heterogeneity analysis. Identity reconstruction quarantines 45 ambiguous groups involving 99 possible users and checks available history-prefix mappings. Missing original ada-002 artifacts and generation logs limit claims about complete historical identity recovery.

D Fresh-MIND Protocol and Complete Results

Reward isolation. The LSA encoder, ordered GRU/SASRec models, and base/augmented ranking heads are trained on the prior cohort before the fresh cohort is scored. The original fresh-MIND run has one frozen configuration and seed per backbone. It uses five fixed folds, four router families, and a 30% budget. N denotes the total OOF cohort; router training uses approximately four fifths of it. The frozen analysis record lists 2,000 bootstrap draws, 50 repetitions per refitted placebo family, and ten matched positive-control repetitions.

Figure 3C shows all eight policy estimates with their perturbation references; Table 7 lists the same values. Appendix D.2 varies router-training size on a fixed test cohort.

Null and positive controls. The 6 prefixes $\times$ 2 backbones $\times$ 4 routers $\times$ 2 placebo types yield 96 reported null summaries. All have intervals containing zero; the largest absolute mean is 0.00168. These overlapping summaries are not 96 independent studies and do not establish calibrated family-wise error. The subsequent permutation follow-up increases each reference family to 1,000 repetitions. Its train-on-disrupted, credit-on-real construction requires a separate validity argument before its

Table 7: Full fresh-cohort policy family ($N = 4954$, $b = 0.30$). The last two columns are empirical 97.5th percentiles of refitted reward-permutation and trained-noise policy gains, credited on real held-out rewards. They are not confidence bounds on the real policy. Intervals condition on fitted OOF policies and are not adjusted for eight comparisons.

Backbone	Router	$G_{0.30}$	95% CI	Perm. upper	Noise upper
GRU	Ridge	$+0.00709$	$[+0.00460, +0.00951]$	$+0.00482$	$+0.00390$
GRU	GBDT	$+0.00471$	$[+0.00233, +0.00714]$	$+0.00304$	$+0.00193$
GRU	Extra Trees	$+0.00551$	$[+0.00322, +0.00789]$	$+0.00308$	$+0.00265$
GRU	KMeans-16	$+0.00371$	$[+0.00132, +0.00613]$	$+0.00283$	$+0.00220$
SASRec	Ridge	$+0.00517$	$[+0.00244, +0.00794]$	$+0.00399$	$+0.00344$
SASRec	GBDT	$+0.00117$	$[-0.00151, +0.00384]$	$+0.00200$	$+0.00219$
SASRec	Extra Trees	$+0.00421$	$[+0.00138, +0.00698]$	$+0.00232$	$+0.00260$
SASRec	KMeans-16	$+0.00179$	$[-0.00117, +0.00464]$	$+0.00257$	$+0.00271$

exceedance rates can be interpreted as permutation p-values; increasing the repetition count does not supply that argument.

At the full cohort size, the matched positive-control pipeline reaches at least 80% recovery at reported HTE-SNR 0.10 for all GRU routers and for SASRec ridge, GBDT, and Extra Trees. SASRec KMeans requires 0.20. This recovery criterion concerns a designed heterogeneous signal, not the 80% mean-test power convention in Eq. 6.

Self-credited oracle. The fresh-cohort self-credited oracle gains are 0.04253 for GRU and 0.05565 for SASRec. Because sorting and crediting use the same realized reward, these quantities contain order-statistic selection effects. They are not attainable policy ceilings, and the difference from a learned router does not estimate remaining predictable heterogeneity.

Attrition and eligibility. Forty-six selected rows fail the frozen complete-row eligibility rule: 10/11/9/5/11 across the five folds, or 0.92% overall. Excluded rows have longer histories on average (21.85 versus 15.66). Retaining only eligible rows changes the sparse-history share by 0.19 percentage points, which is a limited distribution check rather than proof that exclusion is ignorable. Results condition on generation-valid users; a production fallback policy would also need to include failures.

The cohort construction retains at least one positive candidate per source impression. This is a label-informed offline slate construction and remains outside the router feature channel. The delivered source/retained-slate mapping, before-call feature allowlist, and overlap checks document the eligible population. Detailed causes for all 46 original generation failures are not recoverable. One impression per user reduces repeated-user dependence, but does not make shared training folds independent replications.

D.1 Inference boundary and refit stability

The permutation follow-up permutes training rewards, refits OOF scores, and credits pooled top-budget selections on the original rewards. Let $S(X, y)$ denote the OOF scores and $g(S, y)$ their credited selection gain. Its reference statistic is $g(S(X, Py), y)$ for permutation P, whereas a full-statistic permutation of $T(X, y) = g(S(X, y), y)$ evaluates $T(X, Py)$. These distributions need not coincide. OOF fitting alone also does not make the pooled selection indicators independent of their credited outcomes: global reward permutation can move an outcome into a training fold used to score that user, and pooled thresholds depend on other fitted scores.

An exact counterexample uses ten users in five identical-feature pairs, with pair members assigned to opposite folds. A one-nearest-neighbor OOF rule predicts the partner's reward and selects the top three scores. Under an exchangeable reward null, enumerating the 945 equally likely pairings in the provided construction yields 75 nominal 5% rejections, or $5/63 \approx 7.94\%$. This disproves OOF fitting as a general calibration argument. It neither estimates the fresh-MIND false-positive rate nor shows that its observed gains are false. Valid permutation inference requires a stated null, admissible transformations, and an appropriately recomputed statistic (Hemerik & Goeman, 2018).

Formal family-wise significance for the eight fresh-MIND policies remains unresolved; historical exceedance rates are retained in the frozen archive as perturbation summaries.

The prospective HTE code uses a different construction: it fits and credits each statistic on the same permuted reward vector. The train/credit mismatch above does not transfer to that calculation. Its null assumptions still need to match the sampling design. Prospective policy intervals likewise condition on the already locked policy; they do not include repeated pipeline training.

Table 8: Fresh-MIND refit stability for all eight policies. Counts give positive point estimates and positive marginal 95% intervals, respectively. The ten outer-split refits and five pipeline conditions reuse the cohort and are sensitivity analyses, not independent replications or multiplicity control.

Backbone	Router	Split mean	Gain+	CI+	Pipeline mean	Gain+	CI+
GRU	Ridge	+0.00667	10/10	10/10	+0.00641	5/5	5/5
GRU	GBDT	+0.00365	10/10	10/10	+0.00571	5/5	5/5
GRU	Extra Trees	+0.00520	10/10	10/10	+0.00615	5/5	5/5
GRU	KMeans-16	+0.00369	10/10	10/10	+0.00442	5/5	4/5
SASRec	Ridge	+0.00475	10/10	10/10	+0.00442	5/5	4/5
SASRec	GBDT	+0.00142	8/10	1/10	+0.00261	5/5	2/5
SASRec	Extra Trees	+0.00344	10/10	9/10	+0.00527	5/5	5/5
SASRec	KMeans-16	+0.00186	10/10	1/10	+0.00339	5/5	3/5

Table 8 reports all eight policies. Five pipeline conditions comprise the preserved canonical run and four full seed refits. A same-seed rerun is not bit-identical to the canonical export; both are retained. Across the six originally positive policies, 60/60 outer-refit estimates and 59/60 intervals are positive; the corresponding pipeline counts are 30/30 and 28/30. These checks support stability of the observed pattern under the tested changes without certifying a family-wise test.

D.2 Training size with a fixed test cohort

The original ladder changes the evaluated cohort as N grows (Figure 7). The fixed-test follow-up instead holds 1,600 cached test users and a 480-user acquisition budget constant, with disjoint router-training sets of 100, 250, 500, 1,000, 2,000, and 3,354 users. All 48 backbone/router/size cells and 16 paired training-size contrasts are retained. These cached test rewards had been inspected previously, so this is a retrospective training-size sensitivity, not a new confirmatory test.

Figure 6 shows the training-size curve on the fixed evaluation cohort; Section 5.2 reads it as evidence against a universal onset threshold.

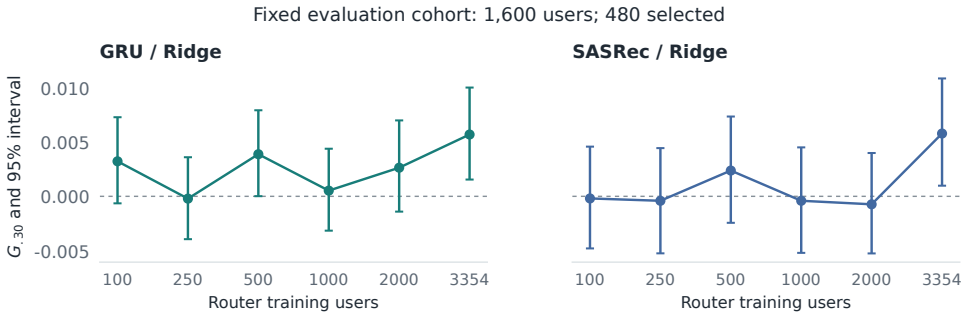


Figure 6: **More router-training users do not yield a resolved monotone gain.** Fresh-MIND ridge selection gain on a fixed 1,600-user evaluation cohort with 480 users selected. Points and 95% fixed-policy intervals use the same evaluation rewards at every training size. Both maximum-size intervals exceed zero, but both 3,354-minus-500 contrasts cross zero.

At 3,354 training users, ridge gains are 0.005693 $[0.001545, 0.010018]$ for GRU and 0.005784 $[0.000980, 0.010857]$ for SASRec. The paired 3,354-minus-500 differences are 0.001814 $[-0.001913, 0.005618]$ and 0.003409 $[-0.001050, 0.008004]$, respectively. Pointwise intervals use

10,000 fixed-policy user bootstrap draws. These comparisons leave the incremental benefit of additional router-training users unresolved at the tested sizes.

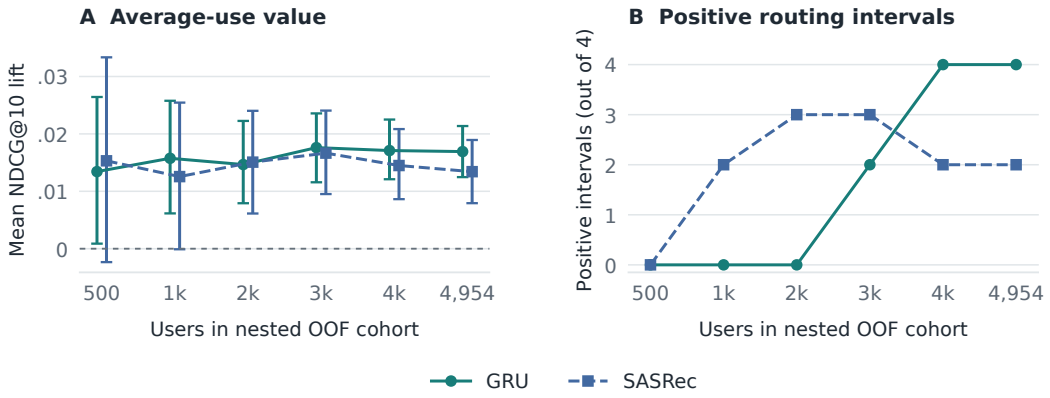


Figure 7: **Historical fresh-MIND scaling is descriptive, not a universal onset threshold. A:** Average-use mean lift with 95% bootstrap intervals. **B:** Number of the four routers with positive marginal policy intervals for each backbone, using the interval criterion alone with no placebo filter. Reward summaries and evaluated users both change across sizes, so the ladder does not isolate the effect of training sample size.

E Prospective Pilot: Frozen Decisions and Independent Outcomes

The prospective pilot separates an audit's forecast from the policy outcome used to assess it. Each source has 800 audit users, 1,600 independent router-training users, and 1,600 evaluation users. GRU and SASRec share users, generated hypotheses, and role assignments within a source. MIND GRU is the sole primary endpoint; the other three conditions are secondary. The Amazon extension adds a second source without changing that designation.

E.1 Eligibility and information boundaries

MIND generation yields 4,005 complete users from 4,100 initial requests, with 95 terminal failures. The first 4,000 complete users populate the roles; five surplus users and all failures remain archived. Amazon generation yields 4,096 complete users and four terminal failures; the first 4,000 populate the roles and 96 remain surplus. Both analyses condition on complete generation. The deployment value of a fallback for failed acquisition is not measured.

The Amazon source is Beauty and Personal Care from Amazon Reviews 2023. Its historical reward models are fitted on All_Beauty before fresh generation. All 631,986 All_Beauty users are excluded from the fresh source. Sampling uses a fixed lowest-5%-hash prefix with 36,350 eligible users; this is the inspected prefix, not a global eligibility census. The task uses next-review relevance proxies, static product metadata, and deterministic hash-selected negatives. MIND retains a positive candidate in each offline slate. These choices define reproducible offline rewards, with a narrower interpretation than label-blind production acquisition.

The final router is ridge regression with regularization parameter 1. Its scaler is fitted on router-training users only. Ten allowed features are available before the call; post-call SHE values and evaluation rewards are excluded. Each model selects 480 of 1,600 evaluation users using the frozen budget and tie handling, then locks the selection mask before the evaluation rewards are scored. Pre/post evaluation features match exactly in the delivered checks. Every final policy is evaluated, including policies whose audit abstains or defers.

E.2 GENERATION AND ROUTER CONFIGURATION

Both prospective sources request the deployment alias `Shared-gpt-5.5-paygo`, using Scheme B, high reasoning effort, and a 1,536-token completion cap per request. The configuration allows at most two requests per user, 300 retries shared across the cohort, and 4,400 total requests for 4,100 initial users. The delivered configuration and launch records specify the requested service; the exact served-model snapshot and billing rates are unavailable. Generated hypotheses are cached and shared across backbones and audit methods. These runs therefore evaluate one requested generator configuration, not robustness across LLM families.

The router uses the following ten-field allowlist:

> `n_candidates`, `n_history`, `history_top1`, `history_margin`, `history_mean`, `history_sd`, `backbone_top1`, `backbone_margin`, `sparse`, `multi_intent`.

History count is capped at 30. History scores are candidate cosine similarities to the normalized mean of history item embeddings; their maximum, top-two gap, mean, and population standard deviation form four features. Backbone scores are standardized within each candidate slate using its mean and population standard deviation, with 10^{-12} added to the denominator; only the maximum and top-two gap enter the router. A single-candidate gap is zero. Sparse and multi-intent flags come from the frozen, source-specific history construction. Amazon's prospective flags use history length at most four and at least three title-derived subcategories, respectively, rather than the original Amazon thresholds. Candidate count uses the slate length, not its labels. The final StandardScaler and ridge model are fitted only on the 1,600 router-training users, with regularization 1 and no hyperparameter search. This input boundary is conditional on the offline slates described above.

E.3 FROZEN RULES

All four data-dependent rules use the same audit users. Mean-only approves when mean reward is positive and its two-sided test has $p < .05$, otherwise deferring. HTE-only requires both permutation statistics to be available; it approves if prediction R^2 or the GATES-style grouped contrast has $p < .025$, otherwise deferring. Each permuted HTE statistic refits and credits the permuted rewards consistently. Missing required evidence is insufficient.

The frozen RAISE-v1 rule is a conservative instantiation of the broader audit: it approves when both component rules approve, defers when both defer, and records insufficient evidence on disagreement or missing evidence. The small cross-validated ridge pilot approves when its audit gain interval lies above $\epsilon = .005$, defers when its upper bound is at most ϵ, and otherwise records insufficient evidence. Always-approve and always-defer provide constant references. Table 9 retains every prediction. In particular, the CV pilot's Amazon GRU decision is an abstention, not an explicit rejection.

Table 9: All prospective decisions, frozen before final policy outcomes. Insufficient denotes an abstention. The combined rule is the tested v1 agreement rule, distinguished from the measurement protocol and the router.

Audit rule	MIND GRU	MIND SASRec	Amazon GRU	Amazon SASRec
Mean-only	Defer	Approve	Approve	Approve
HTE-only	Approve	Defer	Approve	Approve
CV pilot	Insufficient	Insufficient	Insufficient	Approve
RAISE-v1	Insufficient	Insufficient	Approve	Approve
Always approve	Approve	Approve	Approve	Approve
Always defer	Defer	Defer	Defer	Defer

E.4 FREEZE CHRONOLOGY AND ENDPOINT RECONSTRUCTION

Archived event records place the audit freeze before author continuation, independent router fitting, selection locking, and evaluation scoring. Table 10 gives the central boundaries. These timestamps describe artifact chronology; cached replay verifies the numerical consequences of the saved selections, not historical human authorship.

Table 10: Prospective freeze boundaries in UTC, rounded to seconds. The archived records retain finer precision and intermediate author confirmation, router-model, final-outcome, and archive seals.

Event	MIND	Amazon BPC
Audit predictions frozen	Sep 15, 07:22:18	Sep 16, 06:56:24
Evaluation selections locked	Sep 16, 01:32:48	Sep 16, 17:27:49
Evaluation rewards scored/frozen	Sep 16, 01:33:28	Sep 16, 17:29:51

For a locked mask s_i and paired reward Δ_i, the reported estimate is $\widehat{G}_{.30} = 1600^{-1} \sum_i (s_i - .30)\Delta_i$. All four masks have exactly 480 selected users. The 10,000 shared user bootstrap draws resample these centered contributions, conditioning on the fixed models and masks. Resampled selection rates may vary; the frozen reference rate remains .30. The procedure does not refit the router or reselect a new top-budget cohort inside each bootstrap draw. Its intervals are consequently fixed-policy sampling summaries, not refit-aware or distribution-free guarantees.

Numerical reconstruction uses cached paired rewards, selections, seed 20260917, and the shared index manifest. All four gains, interval endpoints, and bootstrap replicate arrays match the delivered exports. The index-matrix SHA-256 is recorded in `data/followup_20260916.json`. No new generation, reward-model training, or outcome-dependent reselection enters this replay.

Table 11: Complete prospective ridge outcomes in absolute NDCG@10 units. The target $\epsilon = .005$, primary designation, and classifications are unchanged from the frozen evaluation. Intervals are marginal and conditional on the fitted policy, not a four-endpoint family-wise claim.

Setting	Role	$G_{.30}$	95% interval	At $\epsilon = .005$
MIND / GRU	Primary	$+0.005427$	$[+0.001288, +0.009739]$	unresolved
MIND / SASRec	Secondary	$+0.004966$	$[+0.000173, +0.009863]$	unresolved
Amazon BPC / GRU	Secondary	$+0.015871$	$[+0.007619, +0.024372]$	confirmed-useful
Amazon BPC / SASRec	Secondary	$+0.017183$	$[+0.009373, +0.025278]$	confirmed-useful

In Table 11, both MIND intervals exceed zero but cross the target. They support positive selection under the reported fixed-policy analysis and leave target attainment unresolved. Both Amazon intervals exceed the target. Because RAISE-v1, mean-only, HTE-only, and always-approve all approve both Amazon conditions, these outcomes do not distinguish their predictive value. There are two correlated confirmed-useful secondary conditions and zero resolved below-target conditions. The latter denominator is unavailable for evaluating false-approval discrimination. Unresolved MIND outcomes are neither successes nor errors for the target-based audit comparison. These literal denominators prevent a four-condition pilot from being interpreted as framework-wide calibration.

F Mean-Power Planning and Sensitivity

Let $\sigma_\Delta > 0$. Under a normal approximation, the standardized sample mean is approximately normal with noncentrality $\sqrt{N}\mu_\Delta/\sigma_\Delta$. A two-sided level-$\alpha$ test rejects when $|Z| > z_{1-\alpha/2}$. For a positive alternative, the usual planning approximation neglects the small opposite tail and solves $\sqrt{N}\rho \approx z_{1-\alpha/2} + z_{1-\beta}$ for target power $1-\beta$. With two-sided $\alpha = .05$ and 80% planned power this yields the mean-effect planning reference quoted in Section 2.2,

$$\rho^\star(N) = \frac{z_{1-\alpha/2} + z_{1-\beta}}{\sqrt{N}} \approx \frac{2.8}{\sqrt{N}}, \qquad N_{\min} \approx \left(\frac{2.8}{\rho}\right)^2. \tag{6}$$

Estimated effects, variance, and dependence introduce uncertainty into the plug-in $N_{\min}$. Reaching $N_{\min}$ is neither necessary nor sufficient for useful routing: it concerns the mean alone, and Section 2 shows that the mean does not determine selection gain.

In Table 12, original MIND and Amazon-Beauty remain below all listed references. REES46 is near the boundary and changes classification as the convention changes; its legacy estimated mean is negative. That sign is relevant to acquire-all, without ruling out a separately validated positive

Table 12: Sensitivity of $\rho/\rho^\star$ to the planning convention. Original cohorts use legacy mean-SNRs 0.048, 0.014, and 0.138 with $N = 1263$, 650, and 498. Fresh GRU/SASRec use $0.1030/0.0674$ at $N = 4954$. These screen mean detectability only; the independent heterogeneity and held-out policy analyses provide the selective-acquisition evidence.

(α, power)	Constant	MIND old	Amazon	REES46	Fresh GRU	Fresh SAS
$(0.10, 0.80)$	2.486	0.69	0.14	1.24	2.92	1.91
$(0.05, 0.80)$	2.802	0.61	0.13	1.10	2.59	1.69
$(0.10, 0.90)$	2.926	0.58	0.12	1.05	2.48	1.62
$(0.05, 0.90)$	3.242	0.53	0.11	0.95	2.24	1.46
$(0.01, 0.80)$	3.417	0.50	0.10	0.90	2.12	1.39

subgroup. The full fresh MIND means exceed each displayed planning reference. This stability does not change their policy intervals or establish routing through the mean screen.

At $N = 500$, fresh GRU has $\rho = 0.0917$ and $\rho/\rho^\star = 0.732$, while its mean-lift interval is positive. There is no inconsistency: the normal rejection reference is approximately $1.96/\sqrt{N}$, whereas $2.8/\sqrt{N}$ plans for 80% rejection probability over repeated samples. Clearing the first in one sample need not clear the second.

G Cost Interpretation of the Fresh-Cohort Gains

For fixed $0 < b < 1$, write $V_b = \mathbb{E}[\pi_b \Delta] = b\mu_\Delta + G_b$. Under a uniform reward-equivalent call cost $\lambda \geq 0$ and no routing overhead,

$$U(\pi_b) = V_b - b\lambda, \qquad U(\text{all}) = \mu_\Delta - \lambda, \qquad U(\text{skip}) = 0.$$

The fitted policy beats both all and skip when

$$\frac{\mu_\Delta - V_b}{1 - b} < \lambda < \frac{V_b}{b}. \tag{7}$$

The interval in Eq. 7 compares fixed policies under an illustrative uniform cost. Appendix M reports the original generation latency that motivates treating the call as costly. The follow-up recomputes value at the exact observed acquisition fraction, $q = 1486/4954 \approx 0.29996$, rather than rounding it to 0.30 (Table 13).

Table 13: Fresh-MIND ridge incremental value retained at the actual acquisition rate. Intervals jointly resample selected and acquire-all rewards over 10,000 user draws, conditional on fitted policies. Retained value is V_q/μ_Δ; it includes the $q\mu_\Delta$ random-acquisition component and is not a fraction of total ranking accuracy or a measured monetary saving.

Ridge	Actual q	G_q	V_q	V_q/μ_Δ	Joint 95% interval
GRU	0.299960	0.00709	0.01216	71.86%	$[58.30\%, 89.22\%]$
SASRec	0.299960	0.00517	0.00919	68.51%	$[48.59\%, 99.12\%]$

The cached budget study retains 48 policy/budget cells and 384 uniform-cost evaluations, including 100 negative utilities. Budgets are 10%, 20%, 30%, 50%, 75%, and 100%; costs range from 0 to 0.05 in reward-equivalent units. At 100% acquisition, selection gain is exactly zero. These are offline value calculations with zero assumed routing overhead, not a learned, priced deployment frontier. Nonuniform token costs, latency, overhead, and failure fallback require additional measurements.

Prospective generation and audit accounting. MIND records 4,302 requests (4,100 initial and 202 retries), 5,755,211 known tokens, and 123 requests without token usage. Amazon records 4,149 requests (4,100 initial and 49 retries), 5,101,068 known tokens, and two requests without usage. Shared generation is counted once across backbones and audit methods; reasoning tokens already included in completion usage are not added again. Missing usage makes known totals lower bounds. The standalone full audit takes approximately 5.59 seconds on MIND and 6.01 seconds on Amazon,

versus roughly 0.077 seconds for the small CV pilot and 0.0006 seconds for the mean-only rule in the recorded environment. These timings exclude generation, reward-model training, and production serving. They do not establish end-to-end latency or dollar savings.

H Structured Hypothesis Embeddings and Average-Use Evidence

SHE provides a concrete observation whose content can be inspected separately from its acquisition value. Figure 8 shows how a history becomes ranking features. The original analyses below examine citation alignment, confidence calibration, and average predictive gains across backbones. Their cohorts are distinct from the fresh-MIND routing study.

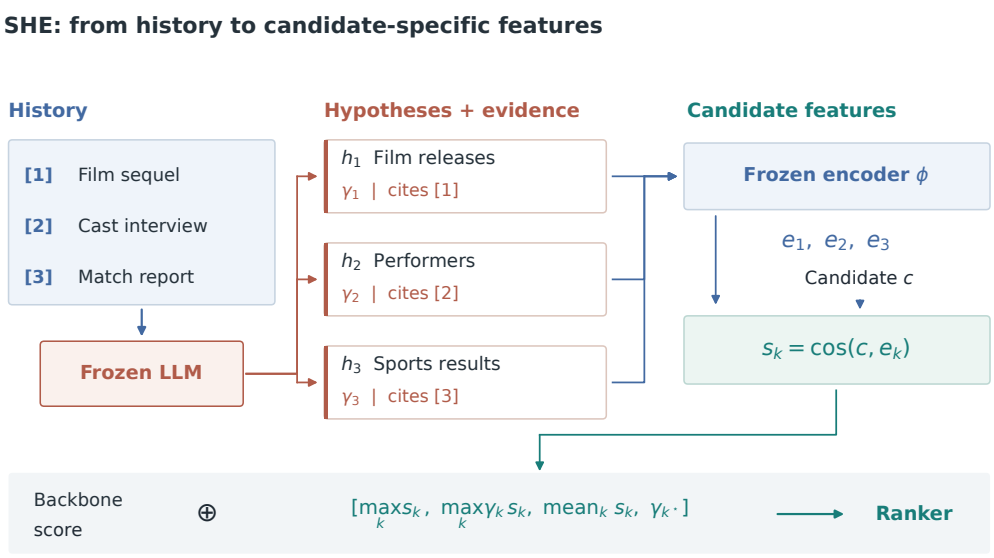


Figure 8: **SHE maps a history to four candidate-specific ranking features.** A frozen LLM generates three hypotheses with confidence scores and evidence indices. Constructed history and hypothesis examples illustrate the format; confidence scores are symbolic. Frozen embeddings yield candidate–facet matches, whose four-coordinate block is concatenated with the backbone score for ranking. The before-call acquisition decision is shown separately in Figure 1.

Embeddings and feature block. The original content-based MIND and REES46 analyses use `text-embedding-ada-002`, $d = 1536$. Amazon-Beauty and the original ordered-backbone MIND study use a local 256-dimensional TF-IDF plus TruncatedSVD (LSA) representation. Fresh MIND also uses a prior-cohort-fitted LSA representation, frozen before fresh scoring; its configuration and checkpoints are included in the private handoff. Hypotheses and items are embedded in the same space within each analysis, with ℓ_2 normalization. Cosines are not compared across different encoders as if they shared a scale.

The original four-coordinate SHE block, whose confidence-weighted coordinate is Eq. 3, is

$$\left[\max_k \cos(\boldsymbol{c}, \boldsymbol{e}_k), \quad \max_k \gamma_k \cos(\boldsymbol{c}, \boldsymbol{e}_k), \quad \frac{1}{K}\sum_k \cos(\boldsymbol{c}, \boldsymbol{e}_k), \quad \gamma_{k^\star}\right],$$

where $k^\star$ maximizes the unweighted facet match. This block is concatenated with the base score. The original MIND late-fusion head is an ℓ_2-regularized, class-balanced logistic ranker with $C = 1$; standardization uses training-fold statistics. The original REES46 study uses a logistic/ridge probe and PCA reduction for its low-N setting. Downstream gradients do not update the LLM or text encoder.

H.1 SEMANTIC CHECKS

On original MIND, hypothesis embeddings are closer to cited than non-cited evidence by 0.0705 cosine points, with 95% CI $[0.068, 0.073]$. The corresponding alignment is approximately zero on REES46, whose windows are 87.9% single-category with thin text. Pairwise hypothesis distinctiveness $1 - \overline{\cos}$ is 0.204 on MIND and 0.104 on REES46. These embedding-space checks make citation support inspectable; they are not human-verified semantic truth labels.

On REES46, five-fold cross-fit isotonic calibration reduces ECE from 0.142 to 0.031 for the cited-evidence-alignment target; the raw Brier score is 0.166. Calibration here concerns that target, not the probability that acquisition improves ranking. Scheme A and Scheme B use different generation settings, so their comparison is descriptive rather than a structure-only ablation. Figure 9 collects these three semantic checks.

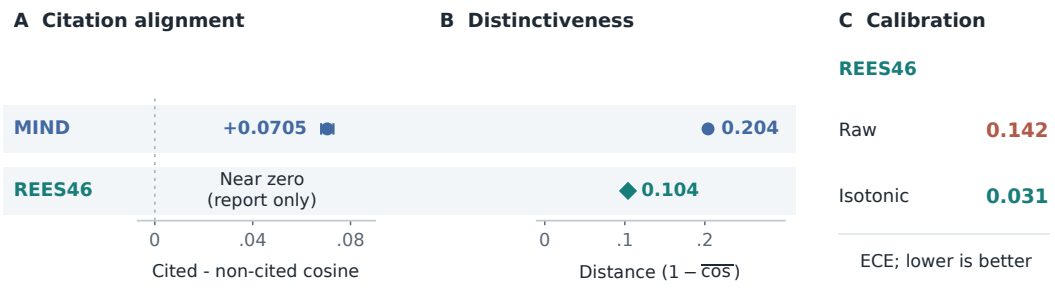


Figure 9: **Original-cohort semantic checks inspect signal content, not selective value. A:** MIND citation alignment: mean cited-minus-noncited cosine difference 0.0705, with 95% CI $[0.068, 0.073]$; REES46 is reported only as near zero. **B:** Mean inter-hypothesis distinctiveness (intervals unavailable). **C:** REES46 cited-alignment ECE before and after five-fold cross-fit isotonic calibration; only aggregate values are available, without bin coordinates or intervals.

H.2 BACKBONE-CONDITIONAL AVERAGE VALUE

Across the original MIND feature ladder, SHE's added mean NDCG@10 lift decreases from 0.0161 at L0 to 0.0146, 0.0100, and 0.0094 as the baseline becomes richer. At L3 its interval is $[-0.0013, +0.0199]$ (Figure 10). This is an average-use redundancy pattern within that analysis, not a routing threshold.

The ordered/unordered 2×2 experiment on 1,411 LSA MIND windows asks a more specific question. The base mean-pool and GRU scores are 0.3701 and 0.3992 NDCG@10. Adding SHE yields $+0.0109$ $[-0.0046, +0.0277]$ over mean-pool and $+0.0114$ $[+0.0030, +0.0209]$ over the ordered GRU. The interaction is -0.0005 $[-0.0164, +0.0150]$, $p = 0.919$ (Figure 11). Thus the data do not establish a different SHE effect between these two backbones even though one individual interval is positive and the other crosses zero.

Original SASRec robustness shows $+0.0179$ over its ordered backbone. On Amazon-Beauty (median history length five), the evaluated GRU score 0.4104 is below mean-pool's 0.4777; ordering alone need not make this fitted backbone stronger. These observations motivate distinguishing sequence availability, backbone strength, and effect size. They are exploratory signal studies rather than evidence that sequential versus nonsequential data universally determine the benefit of SHE.

Exploratory coarse gate. The older sparse/multi-intent gate acquires on 86.4% of original MIND windows. Its contrast with acquire-all is approximately $+0.0001$ with 95% CI $[-0.0038, +0.0042]$. This interval permits gains and losses and does not establish equivalence. The gate was derived from the same data and needs independent validation. Fresh MIND shows that a coarse gate is not the only plausible deployment policy: per-instance routers can also recover value in an appropriate setting.

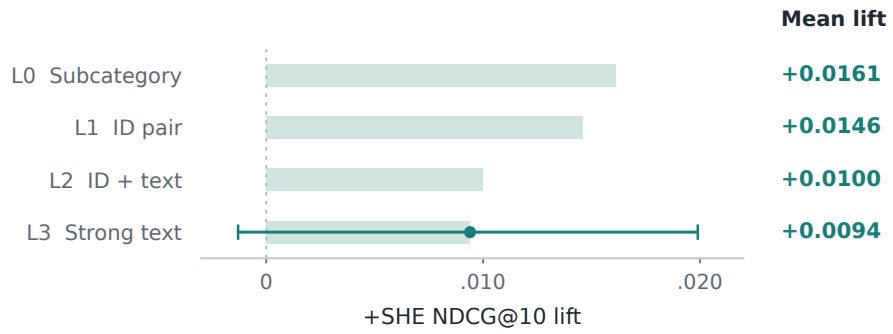


Figure 10: **SHE's average-use point estimates decrease as the original-MIND baseline becomes richer.** Bars show mean SHE gains across four nested feature sets. Only L3 has a numerical 95% CI in the retained summaries; the other bars show point estimates without intervals. This cohort and reward pipeline differ from fresh MIND and do not establish selective value.

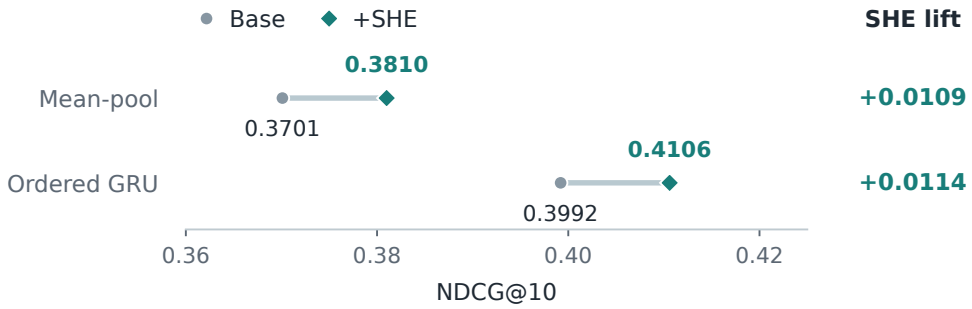


Figure 11: **Positive SHE point estimates do not establish a backbone-ordering interaction.** In the original-MIND LSA study, connected points show each backbone before and after adding SHE. The mean-pool minus GRU lift difference is -0.0005 ($p = 0.919$); the interaction remains unresolved.

I Raw Input Examples

Figure 12 shows the input histories supplied to the frozen LLM in each original dataset; Appendix J reproduces the verbatim generation prompts.

J SHE Prompts and Generation Settings

The original study uses fixed zero-shot prompts to turn histories into structured intent hypotheses. HypoGeneAgent (Yuan et al., 2025) uses a related hypothesis, confidence, and evidence format for gene-set annotation; SHE applies that format to behavioral histories. The prompts request confidence scores, whose empirical calibration is evaluated separately. The fresh-MIND generation manifest remains part of the artifact handoff.

Decoding and model configuration. Scheme A returns one blended-intent summary; Scheme B returns the ranked, evidence-indexed hypotheses used by SHE. Table 14 lists the reported original settings. Both use the same API transport, without fine-tuning, retrieval, or tool use. Their different model versions and reasoning effort make the comparison a joint configuration contrast.

Input serialization. Numbered pre-target events form the history input. MIND uses clicked article titles and categories. E-commerce events supply actions and available category or item descriptions. The step index `[i]` anchors each returned `evidence_indices` entry. The excerpts in Appendix I illustrate the available content; evaluation targets are kept separate from the history prompt.

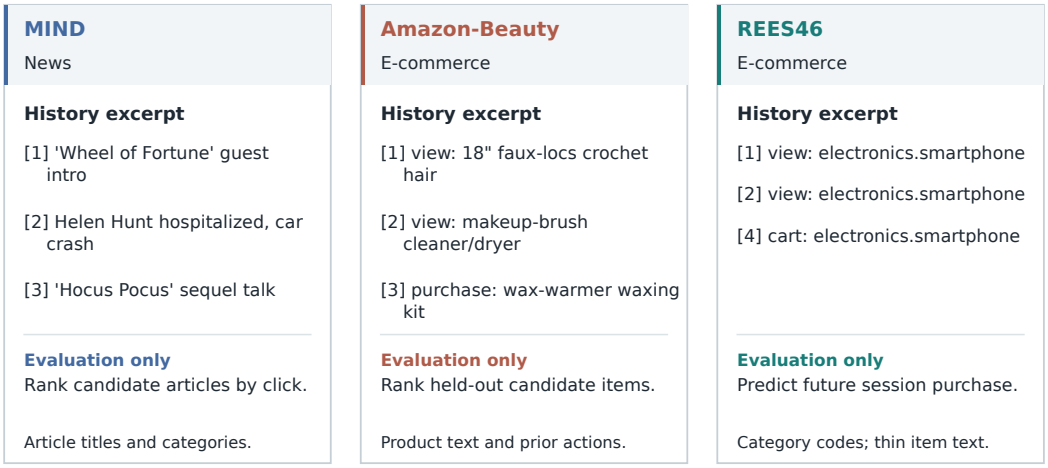


Figure 12: **Original-dataset histories differ in semantic richness.** Abbreviated excerpts illustrate title-rich MIND and Amazon-Beauty histories and a category-code-dominated REES46 window. The lower row describes the evaluation task, separately from the pre-target history supplied to the LLM. These qualitative excerpts provide context rather than performance evidence.

Table 14: Generation settings for the two hypothesis schemes. Scheme B is the source of the SHE feature block (App. H); Scheme A is a descriptive summary baseline, not a matched structure-only ablation.

	Scheme A (summary)	**Scheme B (SHE, Top-3)**
Model	GPT-5.4	GPT-5.5
Reasoning effort	low	high
Output	`{"summary": ...}`	`{"hypotheses": [...]}`
# hypotheses	1 (free sentence)	exactly 3, rank-ordered
Per-item fields	—	hypothesis, confidence, evidence_indices
Fine-tuned?	no (zero-shot)	no (zero-shot)

Output schema (Scheme B). A single JSON object contains exactly three rank-ordered hypotheses. Each has a `hypothesis` statement, a model-reported `confidence` in $[0, 1]$, and `evidence_indices` pointing to supporting history events. The word *calibrated* in the prompt specifies a request, not an empirical guarantee. The fixed hypothesis slots produce a consistent input to the four scalar SHE features in Appendix H.

Boundary rules. The domain prompts specify weak-signal fallbacks, evidence requirements, and restrictions on unsupported demographic inferences. For weak histories, scores must remain below 0.5 and the third hypothesis uses a fixed browsing fallback. The e-commerce prompt requires an add-to-cart trigger for the top scores to exceed 0.7; the news prompt asks for distinct facets when several topics appear. These are generation instructions. Their joint output is assessed by the semantic checks in Appendix H.1; separate rule ablations would be needed to attribute the results to individual instructions.

Verbatim prompts. The complete system prompts follow. The serialized window (above) is appended as the user turn.

News — Scheme A (single summary).

```
You are a news-recommendation and reading-interest expert. Below is a user's recent [reading
↪ history] (news articles they clicked, oldest to newest). In ONE free-text sentence of at
↪ most 30 words, give a single sharp summary of the user's current core reading interest.
↪ Output STRICTLY a JSON object: {"summary": "<your summary>"} -- no markdown, no extra
↪ explanation.
```

News — Scheme B (SHE, Top-3 ranked hypotheses).

```
You are a top expert in news-recommendation and reader psychology. Below is a user's recent [
↪ reading history] (clicked news articles, oldest to newest). Based strictly on the reading
↪ evidence, infer the user's Top-3 ranked [Reading-Interest Hypotheses] -- the distinct
↪ themes/topics the user is most likely to want to read next.
Output STRICTLY a single JSON object: {"hypotheses": [ ... ]} whose array holds EXACTLY 3
↪ elements, ordered by descending strength. No markdown (no ```json), no extra explanation.
↪ Each element must contain these fixed keys:
1. "hypothesis": a specific reading-interest statement with reasoning about the theme.
2. "confidence": a calibrated probability score between 0.0 and 1.0.
3. "evidence_indices": an integer array of the input step indices (the [n] at the start of
↪ each input line) that support this hypothesis.
[Hard boundary rules]
- Weak-signal fallback: if the history is very short or all one narrow topic, every
↪ confidence must be below 0.5, and the 3rd hypothesis must be exactly "Casual browsing / no
↪  strong topical interest".
- Multi-interest capture: when the history spans several distinct topics, the three
↪ hypotheses MUST cover genuinely different interest facets (do not restate one topic).
- Bias elimination: never fabricate the user's absolute gender, age, or region.
```

E-commerce — Scheme A (single summary).

```
You are an e-commerce consumer-behavior expert. Below is a user's [behavior sequence] within
↪ a single shopping window. In ONE free-text sentence of at most 30 words, give a single
↪ sharp summary of the user's current core shopping intent. Output STRICTLY a JSON object:
↪ {"summary": "<your summary>"} -- no markdown, no extra explanation.
```

E-commerce — Scheme B (SHE, Top-3 ranked hypotheses).

```
You are a top expert in e-commerce consumer behavior and user psychology. Below is a user's [
↪ behavior sequence] within a single shopping window. Based strictly on the concrete
↪ behavioral evidence, infer the user's Top-3 most urgent [Ranked Intent Hypotheses] (
↪ immediate purchase motivations).
Output STRICTLY a single JSON object: {"hypotheses": [ ... ]} whose array holds EXACTLY 3
↪ elements, ordered by descending urgency. No markdown (no ```json), no extra explanation.
↪ Each element must contain these fixed keys:
1. "hypothesis": a specific intent-hypothesis statement with behavioral-motivation reasoning.
2. "confidence": a calibrated probability score between 0.0 and 1.0.
3. "evidence_indices": an integer array of the input step indices (the [n] at the start of
↪ each input line) that support this hypothesis.
[Hard boundary rules]
- Weak-signal fallback: if the input is ALL views (view) with NO add-to-cart, every
↪ confidence must be below 0.5, and the 3rd hypothesis must be exactly "Just browsing / no
↪ clear purchase goal".
- Strong-signal trigger: only when an add-to-cart action is present may the top two
↪ hypotheses exceed 0.7 confidence.
- Bias elimination: never fabricate the user's absolute gender, age, or region from category
↪ names.
```

K ORIGINAL-STUDY WINDOW CONSTRUCTION AND SETTINGS

This section describes window construction for the original signal and backbone studies; the fresh-MIND protocol is in Appendix D. All three datasets are public. A *window* pairs a pre-target user history with an evaluation target, and hypothesis generation uses the history. Table 15 gives the exact construction. A "window" is a point-in-time slice: the history is the observed behavior, and the target is a future click (MIND candidate slate), a future purchase (REES46 session), or a held-out next-item slate (Amazon-Beauty). Cold-start (*sparse*) and *multi-intent* slices are defined by fixed thresholds, not tuned.

Per-analysis N. The number of *rerankable* windows in a given result can be smaller than the generated count (windows with an empty valid slate, or missing a valid hypothesis under a scheme, are dropped). We therefore report the exact N with each result (e.g. MIND $N{=}1263$ for the ada-002 late-fusion study, $N{=}1411$ for the backbone-conditional LSA study; Amazon $N{=}650$; REES46 $N{=}498$). No window is dropped for having a favorable or unfavorable reward under the reported eligibility rules. Verifying the exclusions requires the original row manifests.

Statistical unit. Unless otherwise stated, N denotes evaluation windows/impressions and all reported bootstrap intervals resample at that level. The reward-SNR $N_{\min}$ values should therefore be

Table 15: Window construction per dataset. MIND is built with a cohort of 1,600 impressions at stride 5 (keep 1 of every 5) and a 40-candidate slate cap; hypotheses were generated for 1,557 windows (1,549 valid for both schemes). Amazon-Beauty: 650 windows (222 sparse). REES46: 498 session windows. History length is the number of pre-target actions; the *sparse*/*multi* slices are the cold-start and multi-intent regimes analyzed throughout.

	MIND (news)	**Amazon-Beauty**	**REES46**
Domain	news reading	e-commerce	e-commerce
Target	candidate click	next-item slate	session purchase
Windows generated	1,557	650	498
History length	median 19, mean 23.6	median 5	median $\sim$9
Slate size (cap)	40 candidates	20 candidates	— (binary)
sparse rule	$n_{\text{hist}} \leq 5$	$n_{\text{hist}} \leq 5$	$n_{\text{hist}} \leq 5$
multi rule	≥ 8 distinct cats	≥ 2 distinct subcats	(mostly single-cat)
Content richness	high (multi-topic)	high (titles)	low (87.9% single-cat)

read as point-estimate diagnostics under the window-level independence approximation. If multiple windows share a user, overlap in history, or share unusually many candidate items, the effective sample size may be smaller. A user/session-clustered uncertainty audit is the right confirmatory analysis for deployment; we report the window-level diagnostic and treat clustered effective-N as follow-up uncertainty analysis.

L HYPERPARAMETERS

Table 16 retains the reported original-study hyperparameters. These settings should not be taken as a verified configuration manifest for fresh MIND, whose backbones and heads are fitted on a prior cohort and then frozen.

Table 16: Original-study downstream hyperparameters. The ranker, backbones, and LSA space are fixed a priori; the GRU/SASRec inputs are the frozen LSA item vectors, so only the sequence-combination parameters are learned. All CIs use the same impression-level bootstrap.

Component	**Setting**	**Value**
Late-fusion ranker	model	logistic regression (pointwise)
	regularization C	1.0
	max iterations	2000
	class weighting	balanced
	feature scaling	StandardScaler, fit on train fold only
Evaluation	cross-fit	GroupKFold(5), grouped by impression
	prediction	out-of-fold click probability $\rightarrow$ rank
	metrics	NDCG@10, Recall@10, MRR (per impression)
	confidence interval	bootstrap 95%, 2000 resamples by impression
Item text space (LSA)	encoder	TF-IDF + TruncatedSVD
	dimensions	256 (random_state = 0)
GRU backbone	hidden size	64
	optimizer	Adam, lr 10^{-2}, weight decay 10^{-4}
	epochs	15, cross-fit (OOF)
SASRec backbone	attention	2 heads, 1 layer, max length 50
	hidden size	64
	optimizer	Adam, lr 10^{-3}, weight decay 10^{-4}
	epochs	15, cross-fit (OOF)
Robustness runs	seeds	$\{0, 1, 2, 3, 4\}$

M HYPOTHESIS-GENERATION COST AND LATENCY

The original generation runs illustrate why SHE is a costly observation. Table 17 retains their reported latency measurements. They are historical timings for those generation configurations, not

fresh-MIND cost measurements. Fixed-budget routing compares how effectively the call budget is used; a complete serving-cost account must additionally include embeddings, router inference, retries, and other overhead. Appendix G separates that accounting from reward-equivalent break-even calculations.

Table 17: Reported original-generation latency. Scheme B (the SHE source) uses the high-effort setting and has longer mean call latency. The per-window wall-clock figures are measured end-to-end and include API response time plus orchestration overhead, serialization, rate-limit/retry backoff, and worker idle time, so they need not equal the arithmetic sum of the mean per-call latencies. Amazon-Beauty's 650 windows took 284.8 min on a single worker; MIND's 1,557 windows were generated across 8 parallel workers ($\sim$63 min wall). All downstream analyses use cached hypotheses. Independent reproduction additionally requires the corresponding code, cached inputs, and configuration artifacts.

	Scheme A	**Scheme B**	**Per window (A+B)**
Model	GPT-5.4	GPT-5.5	—
Reasoning effort	low	high	—
Latency / call	$\sim$2–3 s	$\sim$6–17 s	—
MIND	—	—	$\sim$19–23 s/win
Amazon-Beauty	3.0 s (mean)	8.6 s (mean)	26.3 s/win

The exploratory original-cohort gate is summarized in Appendix H.2. It remains a candidate coarse policy requiring independent evaluation, not a conclusion that only design-time gating is possible. Fresh-MIND selective policy results are reported separately in Section 5.2.

N Evidence Provenance and Remaining Checks

The frozen analysis record of the original study round supplies the initial follow-up evidence. The figure builder, `scripts/build_revision_assets.py`, parses its tables, checks their internal arithmetic, and produces vector PDFs and TeX tables. The extracted values, source-table locations, and record checksum are recorded in `data/f1_f4_report_values.json`. This makes the presentation traceable; source-code and raw-output reproduction remain separate checks.

The completed follow-up delivers ten frozen packages with source snapshots, cached outputs, split manifests, model artifacts, analysis assets, and checksums. They document the controlled screen's binary rule, the before-call allowlists, removed post-call fields, retrospective refits, and prospective audit/policy boundaries. `scripts/build_followup_assets.py` verifies the five package hashes used for its exports, reads their CSV/JSON/NPZ files, and generates additional vector figures and TeX tables. The four prospective gains and all 10,000 bootstrap replicates per condition have been reproduced from cached rewards and locked selections; provenance is recorded in `data/followup_20260916.json`. This numerical replay is narrower than rerunning LLM generation, reward-model training, or every scientific refit.

The controlled figure uses the six reported corrected cases. Earlier phase curves and an earlier agentic-router experiment are excluded pending verification of their result tables and feature state. Original oracle and average-use summaries retain their original-cohort scope. Later mechanism and prospective experiments are reported with their own reward construction and sample boundaries.

For the original figures, `data/original_figure_summaries.json` records source locations for the reported calibration errors and feature-ladder means. The current plots show these summaries directly. Bin-level calibration coordinates, sparse-slice estimates, degradation-sweep tables, and most ladder intervals are unavailable in the retained numerical materials, so curves and error bars for those quantities are omitted. The L3 interval is preserved from the original appendix. This replaces hand-entered plot arrays with traceable summaries, without implying raw-data reproduction.

The completed retained-value analysis supplies joint uncertainty from cached rewards. No estimate is reconstructed from a rendered image. The historical sample-size plot retains reported mean intervals and counts; the new fixed-test plot uses the delivered numerical learning curve. All eight fresh-MIND policies and all four prospective conditions remain in the record. Formal permutation and family-wise

inference remains open for the reason in Appendix D.1; its historical labels are not adopted as confirmed significance claims. Code will be released upon acceptance.

O Further Connections to Prior Work

These connections guide policy choice, evaluation, and extensions of the audit.

Active information acquisition. Active feature acquisition selects measurements given currently observed information (Aronsson et al., 2025). EDDI combines a partial variational autoencoder with expected information gain (Ma et al., 2019). Generative surrogate models provide acquisition rewards and state information for sequential feature selection (Li & Oliva, 2021). RAISE can audit such a choice once the observation, downstream predictor, and cost are specified. Our experiments cover one optional observation block; multiple adaptive acquisitions would require history-indexed rewards, costs, and allowed information.

Information acquisition within an agent. ReAct interleaves reasoning with actions that obtain external information (Yao et al., 2023). To audit an acquisition decision within such a system, an adapter would expose the allowed before-call context, the optional observation, the downstream decision rule, and the cost model. The gate sees only the before-call context; the downstream rule receives the observation if acquired. Audit data must separately supply independently credited outcomes for the specified acquire/skip comparison. The audit budget includes paid calls needed to collect this initial evidence. For a sequential extension, the comparison must specify the decision history and continuation policy, since an observation can change later actions and future observation opportunities. Repeated simulator rollouts could estimate branch returns; logged interactions instead require suitable off-policy identification and action support. Training and evaluation must separate trajectories or independent task instances, with uncertainty accounting for within-trajectory dependence. Cost accounting must include delays and subsequent calls. This proposed integration interface specifies the information, reward, and cost records needed for an agent acquisition audit. The empirical studies instantiate the single-step setting through optional SHE acquisition and ranking.

Deferral, rejection, and incremental benefit. Selective classification trades coverage for predictive risk (Geifman & El-Yaniv, 2017). Learning to defer models the downstream decision-maker's performance (Madras et al., 2018); consistent surrogate losses support joint learning of a predictor and a deferral rule (Mozannar & Sontag, 2020). Base-predictor uncertainty need not identify where an additional observation improves the outcome. RAISE measures this incremental benefit conditional on available information; budget-matched random acquisition separates selecting well from using the resource more often.

Model pools, cascades, and the information available to a router. FrugalGPT learns cascades that can inspect intermediate responses (Chen et al., 2024); RouteLLM learns query-based model selection from preference data (Ong et al., 2025). Cascade routing provides a unified optimization view (Dekoninck et al., 2025), while BEST-Route jointly chooses the model and the number of candidate responses (Ding et al., 2025). Self-REF learns confidence tokens that inform downstream routing and rejection (Chuang et al., 2025). An intermediate response can guide a later escalation, with its generation cost included. Each gate uses information available before its own call; the feature allowlist in Protocol 1 records this timing boundary for each decision being evaluated.

Benchmarks, risk constraints, and deployment infrastructure. RouterBench supplies model outcomes for reproducible routing comparisons (Hu et al., 2024). LLMRouterBench broadens the model and task coverage and reports both simple-baseline competitiveness and remaining oracle gaps (Li et al., 2026). LLMRouter provides modular routing components and evaluation across additional settings, including personalized and multi-turn tasks (Feng et al., 2026). These resources could support wider audits. Conformal LLM routing calibrates a gate for the probability that the cheap model is wrong while the expensive reference is correct among cheap-routed queries, under its exchangeability assumptions (Uddin & Bauer, 2026). This conditional violation target differs from average selection gain at a fixed budget. RAISE reports empirical selection gains with setting-specific uncertainty; distribution-free risk control would require an additional calibration guarantee.

Estimating heterogeneity and evaluating a decision rule. Causal forests support heterogeneous-effect estimation and inference (Wager & Athey, 2018); meta-learners adapt supervised learners to treatment-effect estimation (Künzel et al., 2019), and the R-learner separates nuisance estimation from the effect-learning objective (Nie & Wager, 2021). The quality of an effect estimate and the quality of a resulting action rule remain different objectives (Fernández-Loría & Provost, 2022). Empirical welfare maximization and policy learning optimize treatment allocation (Kitagawa & Tetenov, 2018; Athey & Wager, 2021). PAPE evaluates improvement over random allocation with the same treatment rate (Imai & Li, 2023); RATE evaluates the benefit concentration induced by a prioritization score (Yadlowsky et al., 2025). GATES-style grouping helps inspect predictable variation (Chernozhukov et al., 2018), while direct held-out evaluation determines the evidence for an actual acquisition rule. RAISE uses mean and regression-fit diagnostics to characterize the reward channel, and measures a policy's selection gain directly on held-out rewards.

Offline ranking contrasts and off-policy evaluation. With logged bandit feedback, only the chosen action's outcome is observed. Doubly robust policy evaluation combines reward and propensity models (Dudík et al., 2011); counterfactual risk minimization accounts for the variance of propensity-weighted objectives when learning a policy (Swaminathan & Joachims, 2015). Our real-data studies instead score both ranking pipelines against retained relevance labels on the same audited examples. The resulting contrast measures incremental offline ranking performance. Estimating effects on production outcomes from logged interactions would additionally require a causal or off-policy estimator with justified identification assumptions, including adequate action support.

Precision, optimization bias, and null controls. Bokelmann & Lessmann (2024) analyze how noisy uplift evaluation affects precision and how outcome adjustment can reduce required test data. Cawley & Talbot (2010) show that optimizing a noisy model-selection criterion can itself cause overfitting. Bastani et al. (2026) study optimistic model-based policy gains even under seemingly favorable prediction and splitting conditions; inference-aware optimization incorporates downstream evaluation power into policy choice (Bastani et al., 2025). Permutation-based evaluation provides a reference for performance after disrupting learned dependence (Ojala & Garriga, 2010). In RAISE, the original oracle invariance is an algebraic consequence of sorting the same rewards used for credit, whereas refitted perturbations probe a learned selection procedure. A formal permutation test additionally requires a null-invariant transformation and consistent recomputation of its full statistic (Hemerik & Goeman, 2018), with an appropriate sampling unit and comparison family.

Inspectable LLM signals and collapse diagnostics. KAR obtains external knowledge for recommendation (Xi et al., 2024), RLMRec aligns LLM-derived semantic profiles with recommendation representations (Ren et al., 2024), and LLMRec augments interaction graphs and profiles (Wei et al., 2024). Their representation and augmentation settings are distinct from a per-instance acquisition decision. SHE is an inspectable signal within this broader design space; its confidence-scored hypothesis format also connects to HypoGeneAgent in gene-set annotation (Yuan et al., 2025). RAGEN-2 diagnoses a failure of input-dependent reasoning during agentic RL (Wang et al., 2026). In both diagnostic settings, an aggregate success indicator can leave the operationally important dependence unmeasured. For RAISE, that dependence is between before-call context and incremental reward. The empirical evidence comes from the specified SHE cohorts and reward pipelines; the broader signal literature supplies further settings in which to investigate this dependence.