TRACEGATE: RANDOMIZED SUPPORT CALIBRATION FOR LOCAL RETRIEVAL EDITS IN REPEATED AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Should a repeated agent place a previously useful fragment into its next retrieval slate after tool reliability or execution policy changes? TraceGate answers with a randomized support-calibration experiment at the candidate occurrence: forced inclusion and exclusion define the verifier-pass contrast, and task-held-out Horvitz–Thompson bin effects evaluate frozen scores without imputing individual potential outcomes. A signed-clipped AIPW estimate may change slate membership only after overlap, a $\Gamma^\star{=}1.5$ stressed interval, uncertainty, and set-size checks agree; otherwise a development-fixed scorer retains control, while an optional bounded probe can resolve ambiguity. The resulting frontier is sharp: over 15,600 retained opportunities, scores attain 87.4% sign agreement and 1.18 pp MAE, while the no-probe rule promotes 64.0% with 6.1% bin-sign disagreement. Bypassing individual gates raises disagreement to 7.4–9.8%; matched-coverage score-only DR and same-pool CRM/DR yield descriptive rates of 10.9% and 9.2%. These local decisions improve completed work: across 9,750 paired episodes, verifier-confirmed success rises by $+1.8/+2.2/+2.6$ pp over the setting-wise strongest non-causal reranker under tool confounding, policy shift, and correlational traps, with all three paired intervals excluding zero; repository-blocked folds and Qwen2.5 retain positive same-pool point differences. By tying each override to a randomized effect target and explicit support evidence, TraceGate converts logged causal estimates into reliable retrieval reuse without additional probes at its primary operating point.

1 INTRODUCTION

A fragment that repairs one tool failure can become the wrong fragment after tool reliability or execution constraints change. Repeated agents encounter this decision again whenever prior repairs, error signatures, and invariant constraints re-enter retrieval. The consequential question is therefore local and counterfactual: how would verifier success change if this candidate occurrence were forced into, rather than out of, the next slate?

We make that contrast measurable on a shared top-128 pool. Forced inclusion inserts one candidate at its generator rank and evicts the lowest-ranked selected fragment; forced exclusion removes it and backfills from generator order. Randomization over this pair yields task-held-out Horvitz–Thompson effects within development-fixed score bins. It supplies an out-of-task calibration target without imputing both potential outcomes for an occurrence, and it exposes where an observational score has assignment support strong enough to guide a local edit.

TraceGate turns this calibration problem into a retrieval rule. A signed-clipped augmented inverse-propensity-weighted (AIPW) score enters a task-clustered interval stressed at $\Gamma^\star{=}1.5$. Overlap, recorded intervention state, interval sign, and a $k{\le}16$ interaction guard determine whether a candidate is promoted, suppressed, optionally probed, or ordered by the development-fixed similarity+tool-state scorer. This estimator-modular design keeps the support-calibrated edit semantics and replayable decision record stable across nuisance-model refits.

The evidence follows the same target at three scales. At the occurrence level, task-held-out bins measure score calibration and reveal whether each gate changes the selected set in the predicted direction. At the episode level, 9,750 paired runs spanning repository repair, WebArena, and GAIA show $+1.8/+2.2/+2.6$ pp verifier-pass gains over the strongest development-fixed non-causal reranker.

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.

Repository-grouped folds and Qwen2.5 then carry the frozen thresholds to new task partitions and a second model family.

The technical contribution is the connection between a randomized local estimand and an executable support rule whose semantics remain fixed across nuisance-model refits. Capped assignment and task-cluster inference make the calibration target auditable; the four-way operator turns overlap, sensitivity, uncertainty, and set interaction into concrete slate behavior; and matched scorers, same-pool CRM/DR, gate ablations, and completed-task outcomes test that connection end to end.

2 Related Work and Problem Distinction

Repeated-agent memory and learned memory control. Trajectory, reflection, skill, and episodic-memory systems make prior experience available to later decisions (Yao et al., 2023; Shinn et al., 2023; Park et al., 2023; Wang et al., 2024; Zhong et al., 2024; Packer et al., 2023). Recent evidence isolates experience-following as a source of error propagation and misaligned replay (Xiong et al., 2026); AgeMem learns storage, retrieval, update, and deletion as agent actions (Yu et al., 2026), while MemCon learns when and how a backend should retrieve, consolidate, or forget (Jiang et al., 2026). TraceGate complements these memory policies with a randomized support-calibration test for occurrence-level reuse after candidate generation.

Causal and risk-aware retrieval. Dense retrieval, late interaction, cross-encoders, and RAG optimize context content (Lewis et al., 2020; Karpukhin et al., 2020; Khattab & Zaharia, 2020; Nogueira & Cho, 2019). CausalRAG and CDF-RAG add causal structure or feedback (Wang et al., 2025; Khatibi et al., 2025); CF-RAG generates counterfactual queries and arbitrates evidence to escape correlational traps (Qin et al., 2026); CoRM-RAG learns perturbation-robust evidence scores with abstention (Liu et al., 2026). Causal retrieval triggering uses randomized exposure and DR uplift to decide whether an upstream generator should run (Hou et al., 2026). TraceGate evaluates a different statistical object: calibration of the forced include/exclude effect, with overlap, sensitivity, and set-size evidence governing whether that estimate changes membership.

Logged-feedback learning, abstention, and slate effects. DR/OPE, counterfactual risk minimization, and unbiased learning-to-rank estimate value or learn policies from logged exposure (Bang & Robins, 2005; Dudík et al., 2014; Swaminathan & Joachims, 2015; Joachims et al., 2017); slate estimators model reward interactions directly (McInerney et al., 2020). Selective prediction studies risk versus coverage (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017), and high-confidence or baseline-bootstrapped policy improvement constrains weak-support updates (Thomas et al., 2015a;b; Laroche et al., 2019). Policy learning with abstention extends this idea to DR objectives and a learned defer action (Sawarni et al., 2026). TraceGate differs in the decision semantics evaluated here: positive and negative local effects produce asymmetric slate edits, resolvable ambiguity can purchase paired probe evidence, unsupported cases retain a matched scorer, and set interaction imposes an explicit operating envelope. The same-pool CRM/DR learner tests whether those branches add value beyond policy-score optimization.

3 From Randomized Local Effects to Executable Edits

The randomized include/exclude contrast defines the quantity to be calibrated; TraceGate determines when an observational estimate of that quantity may change the retrieved set. Explicit precedence over support, ambiguity, and direction emits promote, probe, suppress, or fallback, while the decision record preserves the evidence required to replay each edit.

3.1 Occurrence-level intervention and calibration target

An *opportunity* is a pre-treatment state S_t in which fragment m appears in the shared top-128 pool. Estimator notation treats S_t as the complete pre-treatment surface, including the recorded intervention field Z_t; the typed record displays Z_t separately to make its availability auditable. On the eligible opportunity population $\mathcal{T}_m$ with $k \leq 16$, define

$$s(m) = \mathbb{E}_{t \in \mathcal{T}_m}[Y_t(1) - Y_t(0)],$$

where $Y_t(a)$ is the verifier outcome after forcing arm $a \in \{0,1\}$ and averaging over the logged distribution of the other slate members. Thus $s(m)$ is a marginal inclusion effect on the calibrated opportunity population. Evaluation uses the randomized, task-held-out bin analogue of this quantity rather than an individual potential-outcome label; paired end-task outcomes separately measure the emitted policy. Shared candidate-conditioned nuisance models produce the score: $\hat{Q}(S_t, \phi(m), a)$ predicts $Y_t(a)$ and $\hat{\pi}_m(S_t) = P(I_t(m){=}1 \mid S_t, \phi(m), m \in C_t)$ predicts inclusion. Here $\phi(m)$ contains the fragment embedding, type, generator rank, and tool-state interaction features. Candidate identity is therefore present even though parameters are shared across fragments. Any estimator that returns the score, interval, and support quantities below can replace this implementation without changing the action rule.

Signed-clipped AIPW ranking estimator. Let $I_t(m) \in \{0,1\}$ indicate whether m is included in R_t. The estimator is

$$\boxed{\hat{s}(m) = \frac{1}{|\mathcal{T}_m|} \sum_{t \in \mathcal{T}_m} \left[\hat{\Delta}_t + \bar{w}_t(m) \, (Y_t - \hat{Q}_t) \right]} \tag{1}$$

where $\hat{\Delta}_t = \hat{Q}(S_t, \phi(m), 1) - \hat{Q}(S_t, \phi(m), 0)$, $\bar{w}_t(m) = \text{clip}(w_t(m), [-C, C])$, and $\hat{Q}_t = \hat{Q}(S_t, \phi(m), I_t(m))$.

The centered DR correction weight is $w_t(m) = \frac{I_t(m) - \hat{\pi}_t}{\max(\hat{\pi}_t(1-\hat{\pi}_t), \pi_{\min}^2)}$, $\hat{\pi}_t = \text{clip}(\hat{\pi}_m(S_t), [\pi_{\min}, 1{-}\pi_{\min}])$ with $\pi_{\min}{=}0.02$ and weight clip $C{=}10$. $\hat{Q}$ is a 2-layer MLP (hidden dim 256, MSE loss), and $\hat{\pi}$ is an XGBoost model (128 trees, max depth 6). Both are fitted on the preceding session window and frozen for the next window. The unclipped counterpart of Equation 1 is the standard AIPW/DR score; signed clipping trades exact double robustness for bounded influence. The $C{=}5/10/20$ sweep changes verifier gain by at most 0.4 pp (Table 10). All paper intervals resample the highest-level task cluster over its seeds, episodes, and co-assigned edits; 200 resamples yield the reported 95% intervals. Figure 3 uses 90% intervals for task-held-out calibration bins.

3.2 Replayable typed decision records

For replay, the storage structure must retain dependencies among pre-treatment state, retrieval decisions, interventions, and outcomes. Each record is implemented as a typed DAG with observations O_t, retrieved set R_t, actions A_t, tool calls/results T_t, interventions Z_t (e.g., tool reliability), and verifier outcome Y. For candidate m, the boundary input is $d_t(m) = (\hat{s}_m, I_m(1), I_m(\Gamma^\star), \text{ESS}_m, \hat{\pi}_m, k, b_t)$, where $\Gamma^\star{=}1.5$ is fixed on development data, b_t is the remaining probe budget, and k is checked against the interaction-calibrated set-size guard (Sec. 3.4). The record binds this evidence, the pre-specified fallback comparator, the triggered gate, and the emitted action to the same local edit. It is the common join key for decision replay, randomized calibration, comparator analysis, and direct-set intervention.

3.3 Identification assumptions and boundary operator

Using the logged score $\hat{s}(m)$ to rank fragment-level marginal effects requires sequential ignorability conditional on pre-treatment S_t, candidate features $\phi(m)$, and recorded intervention variables Z_t, together with positivity across the eligible top-128 opportunity surface. The treatment is forced inclusion of m, the outcome is verifier-confirmed task success, and descendants of inclusion such as tool outputs and verifier results are excluded from S_t. Randomized assignments identify the calibration-bin effects without this ignorability assumption; negative controls and odds-ratio sensitivity intervals stress the observational score used online.

Formal assumptions. TraceGate requires three identification assumptions: (A1) Sequential ignorability: $(Y_t(1), Y_t(0)) \perp I_t(m) \mid S_t, \phi(m), Z_t$. (A2) Positivity: $0 < \pi_{\min} \leq P(I_t(m){=}1 \mid S_t, \phi(m)) \leq 1{-}\pi_{\min}$ for all eligible fragments. (A3) Stationarity: throughout an evaluation window, outcome model Q and propensity π remain unchanged; we hold $\hat{Q}, \hat{\pi}$ fixed per session, re-estimating them at session boundaries from the preceding window's episodes (every 500 episodes). These

are standard longitudinal identification conditions (Hernán & Robins, 2020; Robins et al., 2000); randomized calibration instead uses known assignments.

Overlap statistic. We compute ESS from non-negative clipped inverse propensity weights for include/exclude: $w_t^{(1)} = \text{clip}(I_t(m)/\hat{\pi}_t, C)$, $w_t^{(0)} = \text{clip}((1 - I_t(m))/(1 - \hat{\pi}_t), C)$, and $\text{ESS} = \min\{(\sum_t w_t^{(1)})^2/\sum_t (w_t^{(1)})^2, (\sum_t w_t^{(0)})^2/\sum_t (w_t^{(0)})^2\}$. Weak overlap triggered matched ranking fallback in 5.9% of confounded-error decisions and 9.4% under policy shift.

Four-action evidence gate. For any $\Gamma \geq 1$, the admissible propensity set is

$$\mathcal{P}_t(\Gamma) = \left\{ \tilde{\pi}_t : \Gamma^{-1} \leq \frac{\tilde{\pi}_t/(1-\tilde{\pi}_t)}{\hat{\pi}_t/(1-\hat{\pi}_t)} \leq \Gamma \right\}.$$

Within each task-cluster bootstrap draw, we minimize and maximize Eq. 1 over $\tilde{\pi}_t \in \mathcal{P}_t(\Gamma)$; the 2.5th and 97.5th percentiles of those extrema form $I_m(\Gamma) = [L_m(\Gamma), U_m(\Gamma)]$. Larger Γ can therefore only widen the sensitivity interval. TraceGate evaluates the prespecified $I_m(\Gamma^\star)$ with $\Gamma^\star$=1.5, rather than estimating an edit-specific stress level. Let $G_{\text{sup}}(d_m)$ require ESS$\geq$25, $\min(\hat{\pi}_m, 1 - \hat{\pi}_m)\geq$0.03, recorded Z_t, and $k\leq$16, the preselected interaction-calibrated set envelope. Let $G_{\text{amb}}(d_m)$ hold when $L_m(\Gamma^\star) \leq 0 \leq U_m(\Gamma^\star)$. In the ablations, "without overlap" bypasses only the ESS and arm-propensity clauses, "without sensitivity" substitutes the nominal interval $I_m(1)$ for $I_m(\Gamma^\star)$, and "without uncertainty" routes by the point sign when the interval crosses zero; the recorded-state and set-size clauses remain fixed. TraceGate applies the following precedence:

$$\mathcal{B}(d_m) = \begin{cases} \text{ranking fallback}, & \neg G_{\text{sup}}(d_m), \\ \text{probe}, & G_{\text{amb}}(d_m) \ \wedge \ b_t > 0, \\ \text{ranking fallback}, & G_{\text{amb}}(d_m) \ \wedge \ b_t = 0, \\ \text{promote}, & L_m(\Gamma^\star) > 0, \\ \text{fragment suppress}, & U_m(\Gamma^\star) < 0. \end{cases} \tag{2}$$

Ranking fallback discards the local AIPW score and delegates position to similarity+tool-state; fragment suppression removes a candidate whose stressed interval is negative. A probe forks the saved pre-treatment state into cheap-verifier include and exclude branches and records the surrogate contrast $\Delta_{\text{probe}} \in \{-1, 0, 1\}$. This contrast never enters the Y_t-based AIPW fit or $I_m(\Gamma^\star)$: a development-fixed surrogate rule promotes on $+1$, suppresses on -1, and falls back on 0. Probe eligibility is then closed for that candidate, preventing a second probe; both verifier calls count against the method's budget. TG-DR fixes $b_t = 0$; TG-SLA and TG-full expose bounded and accuracy-envelope probe budgets.

Sensitivity to unobserved confounding. The Rosenbaum parameter Γ bounds the factor by which an unobserved variable can change treatment odds (Rosenbaum, 2002). Table 9 reports the complete sweep; the online branch always uses the development-fixed $\Gamma^\star$=1.5.

Probe rollouts. For ambiguous supported candidates, the paired cheap-verifier branches use compile-only, a small unit-test subset, or an SQL constraint check. The verifier choice and surrogate-to-action mapping are fixed on development data. TG-SLA enables these probes under a 3% rate cap and 1.5 s wall-time cap; TG-DR sets the probe budget to zero; TG-full uses a 25% rate without the per-probe cap to characterize the accuracy/latency envelope. Every branch result and post-probe action is added to the typed decision record.

Algorithm 1 TraceGate evidence-to-action control.

Observe state features S_t and intervention variables Z_t
Retrieve top-128 candidate set C by similarity
for each $m \in C$ **do**
  Estimate $\hat{s}(m)$ and $I_m(\Gamma^\star)$; compute ESS and $\hat{\pi}_t$
  if ESS < 25, arm propensity < 0.03, required Z_t is absent, or $k > 16$ **then**
    emit ranking fallback for m
  else if $0 \in I_m(\Gamma^\star)$ **then**
    if budget permits, probe once and route by the fixed surrogate rule; otherwise emit ranking fallback
  else if $L_m(\Gamma^\star) > 0$ **then**
    emit promote
  else
    emit fragment suppress
  end if
end for
Order promoted candidates by lower bound, then fallback candidates by the fixed scorer; break ties by generator rank and fragment hash
Emit the first k non-suppressed candidates; if needed, backfill from remaining non-suppressed candidates in generator order

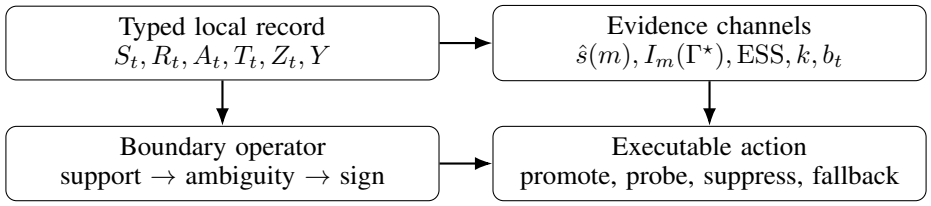


Figure 1: A candidate-conditioned score enters support, stressed-interval, and set-size checks before it can change the slate; every branch records its evidence and action.

Figure 1 separates evidence production from action. Logged pre-treatment state and interventions define the replayable object, modular estimation contributes a signal, and the gate records its action with the matched comparator. Probes are disabled in TG-DR, so ambiguous supported candidates follow the deterministic fallback branch.

3.4 Set-valued retrieval and interference

Retrieval returns a *set* of fragments, so we approximate its propensity with $\hat{\pi}(R_t \mid S_t) \approx \prod_{m \in C_t} \hat{\pi}_m(S_t)^{I_t(m)}[1 - \hat{\pi}_m(S_t)]^{1-I_t(m)}$, where C_t is the shared top-128 pool. This factorization assumes conditionally independent inclusion decisions and is stress-tested under fragment interactions (Tchetgen Tchetgen et al., 2021). On a randomized slate slice, a cross-fitted slate-response model supplies the post hoc prediction $\hat{\tau}^{\mathrm{model}}_{\mathrm{slate}}(R)$, and we define the model-based interaction diagnostic $\rho(R) = \left|\hat{\tau}^{\mathrm{model}}_{\mathrm{slate}}(R) - \sum_{m \in R} \hat{s}(m)\right|$. Its median/p95/p99 are 0.41/1.32/2.14 pp at k=16, compared with 0.28/0.94/1.61 at k=8 and 0.63/1.78/3.02 at k=32 (Table 7). These residuals set a conservative score-driven envelope at $k \leq 16$, with larger sets delegated to matched fallback; paired end-task comparisons evaluate the selected operating point.

4 Operating Points and Cost Model

TG-DR is the default no-probe operating point: top-128 candidates from bge-large-en-v1.5 with FAISS HNSW, $\pi_{\min}$=0.02, ESS threshold 25, fixed $\Gamma^\star$=1.5, interaction-calibrated $k \leq 16$, and deterministic similarity+tool-state fallback. TG-SLA adds probes at at most 3%, caps each paired probe at 1.5 s, and counts any episode exceeding 8.0 s as a failure. TG-full uses a 25% probe

rate without the per-probe cap to characterize the accuracy/latency envelope. All configurations share Eq. 2; probe availability alone changes b_t. TG-DR isolates the action rule from probe-time computation.

DR scoring adds <15 ms p95 per step. Weak overlap triggers ranking fallback rather than fragment removal, while a negative supported interval triggers fragment suppression. The observed 5.9% and 9.4% overlap-fallback rates are measured properties of the confounded and shift settings. Runtime decomposition appears in Appendix G.

5 Experimental Design

The experiments follow three questions in causal order: **RQ1**, do frozen local scores align with randomized include/exclude effects where TraceGate promotes; **RQ2**, do the support checks improve both promoted-edit calibration and verifier-confirmed task completion against matched scorers; and **RQ3**, do development-fixed thresholds retain those properties under task/repository blocking and a second model family?

Estimand, assignment, and inference unit. The top-8 audit population is induced by the shared top-128 generator on 650 recurring tasks. For an opportunity $o = (t, m)$, the intervention replaces only m's membership in the current slate: forced include inserts m at its generator rank and evicts the lowest-ranked selected fragment; forced exclude removes m and backfills the highest-ranked unselected fragment. Generator rank followed by fragment hash resolves every tie. The randomized estimand is the include-minus-exclude verifier-pass difference within a prediction bin, averaged over the design distribution of co-edits. The task is the highest-level inference cluster: all seeds, episodes, opportunities, and co-assigned edits for that task are resampled together.

Evaluation units and aggregation. Each benchmark task under one RNG seed is an episode paired across methods. We report verifier-pass rate averaged over five independent seeds $\{13, 37, 61, 71, 97\}$. Seed standard deviations summarize run variation; paired 95% difference intervals use a task-cluster bootstrap over the five-seed task mean; paired permutation tests use the same 650 task units with α=0.05. Models are frozen inside 500-episode stationarity windows.

Hardware, models, and workloads. Primary experiments ran on 4$\times$ NVIDIA A100 80 GB GPUs with Llama-3.1-70B-Instruct in vLLM 0.5.4 at FP16 (CUDA 12.4, PyTorch 2.3.1, driver 550.54.15). The cross-family replication uses Qwen2.5-72B-Instruct (Yang et al., 2024) on the same hardware and precision; prompts, tool versions, fragment pool, candidate generator, thresholds, and seeds are fixed, while model-dependent nuisance functions and learned comparator policies are refit on Qwen preceding-window logs. Embeddings use `bge-large-en-v1.5` v2.0 with FAISS HNSW (M=32, ef=128). The 650-task suite contains 300 issue-level repository-repair tasks from 54 public repositories, encoded in the SWE-bench instance schema and evaluated with per-instance patch tests (Jimenez et al., 2024); this custom collection is not the official SWE-bench Verified split. The remaining tasks are 200 WebArena v0.4 tasks (Zhou et al., 2024) with deterministic graders and 150 GAIA v2025-03 tasks (Mialon et al., 2024) with exact-match plus rubric grading.

Training and evaluation separation. The primary protocol fits $\hat{Q}$ and $\hat{\pi}$ on the immediately preceding 500 episodes for the same model and setting, freezes them, and scores the next session. A chronologically earlier warm-start window supplies the first fit and is excluded from the 9,750 reported episodes. Thresholds are selected once on a separate controlled-family development partition by minimizing bin-sign disagreement subject to at least 60% promoted coverage; this partition is disjoint from primary, blocked, and Qwen scoring. The blocked analysis scores every task once in disjoint held-out folds, groups all repository-repair tasks by repository, and prevents both nuisance fitting and same-pool learning from seeing the evaluated task or repository.

Baselines and matching. All methods share the base LLM, prompts, tool versions, top-128 candidate generator, fragment pool, and evaluation harness; only retrieval scoring changes. Matched-budget comparisons enforce identical probe-rate, wall-time, verifier-call, and pre-treatment feature access. The nearest logged-feedback alternative is a same-pool CRM/DR learner trained on the same windows, candidate pool, and state surface. The non-causal set contains: (1) Sim: `bge-large-en-v1.5` + HNSW over the same top-128 candidate set. (2) Cross-encoder reranker: `bge-reranker-v2-m3` (Chen et al., 2024; Nogueira & Cho, 2019) over that top-128 set (max length 512, FP16). (3) Learned MLP: an MLP with 2 layers and hidden dimension 256 takes candi-

Table 1: Separation of selection, fitting, and scoring.

Stage	Data access	Frozen object
Threshold selection	separate development partition	$\pi_{\min}$, ESS, $\Gamma^\star$, probe caps, frontier cutoffs
Warm start	chronologically earlier 500 episodes	first $\hat{Q}, \hat{\pi}$, same-pool learner
Window scoring	subsequent 500 episodes	thresholds and preceding-window models
Blocked/Qwen scoring	disjoint fold or model family	thresholds, bins, comparator identities

date embeddings together with tool-state features; training uses preceding-session logs, with neither DR correction nor gates. (4) LLM-judge reranker: relevance scoring with Llama-3.1-8B-Instruct uses batches in FP16 with a maximum of 256 tokens. (5) History-aware router: recent tool calls, retrieval outcomes, and state features feed a Transformer with 4 layers, hidden dimension 256, and context window 8. The full recipes and additional baselines (similarity+tool-state, stratified retrieval, randomized+IPS) are given in Appendices B and I. Randomized+IPS is an exposure-evaluation control. Main results report the development-fixed strongest non-causal method and the same-pool learner separately, avoiding post-test comparator selection.

Controlled settings. The interventions Z_t isolate three recurrent failure mechanisms. *Confounded tool errors* reverses the reliable/unreliable tool assignment at the session midpoint; *policy shift* replaces the active execution-constraint set at a window boundary; *trap stress* inserts a lexically similar but causally irrelevant pair in a preselected 25% of eligible retrieval opportunities. The intervention schedule is fixed before method execution and shared across paired seeds. Fixed episode counts and 500-episode windows preserve pairing, while zero, bounded, and 25% probe budgets expose the utility/latency trade-off.

Table 2: Evaluation protocol. Each setting contains 3,250 paired episodes, freezes nuisance models inside 500-episode windows, and varies only the configured probe budget.

Setting	Episodes	Windows	Intervention	Probes	Purpose
Confounded tool errors	3,250	6×500 + 250	reliability flip	0 / 3 / 25%	induce logged confounding
Policy shift	3,250	6×500 + 250	constraint change	0 / 3 / 25%	test nonstationarity
Trap stress (25%)	3,250	6×500 + 250	trap injection	0 / 3 / 25%	decouple similarity from utility

Randomized validation protocol. Each setting contains 650 tasks × 5 seeds. Before assignment, the frozen generator and policy snapshot materialize a complete top-8 opportunity frame from saved pre-treatment states; intervention outcomes cannot change that frame. Opportunities are masked independently with probability 0.15, a logged uniform permutation caps retention at three per episode, and a fair coin assigns forced include or exclude. The exact cap-adjusted probability $q_e/2$, Horvitz–Thompson bin estimator, and treatment of unretained opportunities appear in Appendix A. Task-cluster resampling preserves dependence from the cap, co-edits, and shared outcomes. Pooled across settings, 183,360 opportunities produce 27,504 pre-cap masks, 7,936 intervened episodes, and 5,200 retained edits per setting (15,600 total): 7,812 include and 7,788 exclude assignments.

Calibration label and counts. Potential outcomes are not imputed per occurrence. Within each setting, equal-mass deciles of the frozen TraceGate score $\hat{s}$ are fixed on development data. For every target task cluster c and bin g, the Appendix A estimator uses randomized assignments in all other task clusters. Each occurrence inherits only this task-held-out bin effect for TraceGate calibration. Sign agreement is $\sum_o \mathbf{1}\{\operatorname{sign}\hat{s}_o = \operatorname{sign}\hat{\tau}^{(-c(o))}_{g(o)}\}/15{,}600$; MAE averages $|\hat{s}_o - \hat{\tau}^{(-c(o))}_{g(o)}|$. The resulting edit-weighted counts are 13,634 agreements and 1,966 disagreements (87.4%). Coverage tables apply each method's development-frozen selection rule to the same occurrences and common

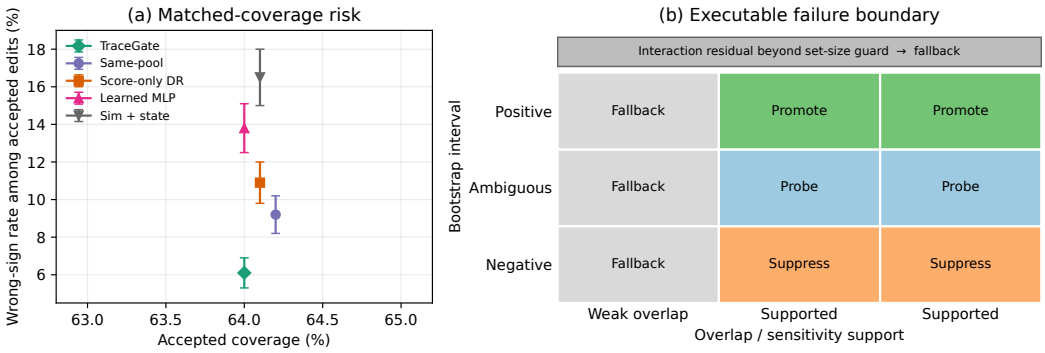


Figure 2: Support calibration at the local-edit operating point. Left: disagreement with task-held-out TraceGate-score-bin signs at development-frozen matched coverage; intervals cluster by task, and the horizontal axis is each method's selected set. Right: overlap, sensitivity, uncertainty, and the k=16 guard map evidence to the four TraceGate outcomes.

TraceGate-score-bin labels. Those cross-method rows are descriptive common-reference diagnostics rather than symmetric calibration comparisons; the paired episode analysis supplies the policy-level endpoint.

Threshold and operating-point selection. The reported operating point is fixed once on the controlled-family development partition. The at-least-60% coverage constraint selects TraceGate's gate thresholds; the "frontier cutoffs" in Table 1 only match comparator coverage for reporting and do not enter Equation 2. Table 1 then separates preceding-window nuisance fitting from subsequent-session scoring. The same gate thresholds, calibration bins, and comparator cutoffs are carried unchanged into repository-grouped blocked folds and Qwen2.5 cross-family evaluation; model-dependent nuisance functions and learned comparator policies are refit. This design tests the decision rule after fixation on two distinct generalization surfaces.

6 Results

Results follow the intervention chain: randomized local effects establish the calibration target, gate ablations test the support mechanism, paired episodes measure completed-task utility, and blocked and cross-family runs test threshold fixation.

6.1 Randomized effects calibrate promoted edits

Each setting contributes 5,200 retained local interventions. Against task-held-out randomized-bin effects, TraceGate attains 87.4% $\pm$ 2.1% sign agreement and 1.18 $\pm$ 0.14 pp MAE over 15,600 occurrences; both "$\pm$" values are 95% task-cluster bootstrap half-widths. Frozen predictions track the $[-3, +6]$ pp region, with most disagreements concentrated at low ESS and high variability.

TG-DR promotes 9,978 of 15,600 occurrences (64.0%); 604 disagree with their held-out TraceGate-score-bin sign, a 6.1% rate. At matched selected coverage, score-only DR and same-pool CRM/DR have descriptive common-reference rates of 10.9% and 9.2%. Removing the overlap, sensitivity, or uncertainty branch changes disagreement to 8.7%, 7.4%, or 9.8% and pooled verifier gain from $+2.2$ pp to $+1.8$/$+1.9$/$+1.5$ pp. The aligned movement in calibration and utility isolates the operational value of the gates beyond the AIPW score.

6.2 Support-calibrated edits improve completed tasks

TG-DR improves over the development-fixed strongest non-causal method by $+1.8$, $+2.2$, and $+2.6$ pp. Paired task-cluster intervals are [0.7, 2.9], [1.0, 3.3], and [1.4, 3.8]; permutation tests give $p = 0.006$, 0.003, and < 0.001. Win/tie/loss counts over 650 tasks are 248/324/78, 265/301/84, and 288/276/86. The no-probe rule therefore improves task success without purchasing extra verifier calls.

Table 3: Verifier-pass rate at k=16 (mean±seed standard deviation). Candidate pool, state access, and budgets are matched; CRM/DR is the nearest logged-feedback learner.

Method	Confounded	Shift	Trap 25%
Similarity	37.8 ± 2.1	40.6 ± 1.9	35.7 ± 2.3
Cross-encoder	41.3 ± 1.5	44.5 ± 1.4	39.2 ± 1.7
Learned MLP	41.6 ± 1.4	44.8 ± 1.3	39.6 ± 1.5
LLM judge	42.4 ± 1.3	45.3 ± 1.2	40.1 ± 1.6
History router	42.1 ± 1.4	45.7 ± 1.3	40.8 ± 1.5
Same-pool CRM/DR	43.5 ± 1.6	47.0 ± 1.5	42.3 ± 1.7
TG-DR, no probes	44.2 ± 1.6	47.9 ± 1.5	43.4 ± 1.7
TG-SLA, 3% probes	44.9 ± 1.7	48.6 ± 1.5	44.1 ± 1.8

The closer same-pool comparison is smaller and setting-dependent. TG-DR leads by 0.7/0.9/1.1 pp; confounded and trap comparisons give $p = 0.041$ and 0.028, while policy shift gives $p = 0.083$. Per-workload confounded effects against the strongest non-causal method are +1.5 [0.2, 2.9] pp on 300 repository-repair tasks, +2.1 [0.4, 3.8] pp on 200 WebArena tasks, and +2.0 [-0.2, 4.2] pp on 150 GAIA tasks.

6.3 Threshold fixation survives model and task blocking

Qwen2.5 evaluation keeps 650 tasks, five seeds, and 3,250 episodes per setting, with prompts, tools, candidates, calibration bins, and thresholds fixed. TG-DR reaches 43.8 ± 1.6/47.0 ± 1.5/42.5 ± 1.8% success, 85.9% cross-fitted sign agreement, 1.27 ± 0.16 pp MAE, and 64.8% promoted coverage, within 0.8 pp of the primary coverage point. Success "±" values are seed standard deviations; the MAE "±" value is a 95% task-cluster bootstrap half-width. Relative to the same-pool learner, point differences are +0.9/+0.9/+1.3 pp (Appendix L).

Repository-grouped blocked folds preserve +0.6/+0.5/+0.7 pp differences over the same-pool learner, with paired intervals [0.1, 1.2], [-0.1, 1.1], and [0.1, 1.4]. Confounded and trap effects exclude zero on both blocked and primary surfaces. Same-feature score ablations reinforce the mechanism: removing propensity, the AIPW correction, or Z_t from nuisance models gives up 2.8, 2.2, and 2.5 pp.

7 Discussion

The randomized opportunity frame and paired policy evaluation answer complementary questions. Task-held-out bin effects test whether frozen local scores have the right direction where the gate promotes, while episode outcomes test whether the resulting slates complete more tasks. Gate removals worsen both bin disagreement and pooled verifier gain, and same-pool CRM/DR narrows the comparison without erasing the confounded-tool and trap advantages. The calibration label is a cluster-held-out group effect, so it supports selected-set calibration rather than an individual potential-outcome claim; the randomized top-8 opportunity frame, $k \leq 16$ guard, and measured interaction residuals define how that evidence enters the emitted slate. Across recurring repository-repair, web, and assistant tasks under controlled tool, policy, and trap shifts, repository blocking and Qwen2.5 retain positive same-pool point differences as nuisance models are refitted; the policy-shift point difference is positive but unresolved because its primary test and blocked interval include zero.

8 Conclusion

TraceGate makes a randomized include/exclude effect operational: local evidence changes a slate only after its support, stressed interval, uncertainty, and set size agree. The resulting no-probe rule combines 6.1% promoted-set bin disagreement at 64.0% coverage with gains over the strongest non-causal reranker in all three repeated-agent stress settings. Randomized support calibration therefore offers a concrete interface between logged causal evidence and reliable retrieval reuse.

Reproducibility Statement

Sections 3.1–5 specify the estimand, assignment, estimator, diagnostics, thresholds, models, hardware, seeds, workloads, session windows, randomization, and task-cluster inference. Appendices B and D–O provide baseline recipes, robustness sweeps, implementation timings, feature access, blocked-fold construction, and complete cross-family results.

Ethics Statement

TraceGate filters retrieval overrides through support checks, sensitivity analysis, and logged fallback decisions. Tool outputs and typed traces can contain sensitive content; deployments should apply access control, retention limits, and redaction, and should restrict traces to retrieval reliability analysis rather than user profiling or surveillance.

AI Use Statement

Generative AI tools assisted with manuscript editing, presentation, formatting, and reference verification during preparation. The authors reviewed AI-assisted edits and take responsibility for the final manuscript.

References

Heejung Bang and James M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-Embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2318–2335, 2024. doi: 10.18653/v1/2024.findings-acl.137.

Miroslav Dudík, Dumitru Erhan, John Langford, and Lihong Li. Doubly robust policy evaluation and optimization. *Statistical Science*, 29(4):485–511, 2014.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4878–4887, 2017.

Miguel A. Hernán and James M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC, 2020.

Junpeng Hou, XianXing Zhang, Sai Xiao, Derek Cheng, Darren Reger, Olafur Gudmundsson, Mehdi Ben Ayed, Zhiqing Rao, and Huizhong Duan. Deep-learning causal retrieval optimization for efficient e-commerce distribution in pinterest. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 7391–7400, 2026. doi: 10.1145/3770855.3818355.

Eric Hanchen Jiang, Zhi Zhang, Yuchen Wu, Levina Li, Dong Liu, Xiao Liang, Rui Sun, Yubei Li, Edward Sun, Haozheng Luo, Zhaolu Kang, Aylin Caliskan, Kai-Wei Chang, and Ying Nian Wu. Memory as a controlled process: Learned adaptive memory management for LLM agents. *arXiv preprint arXiv:2607.13591*, 2026.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R. Narasimhan. SWE-bench: Can language models resolve real-world GitHub issues? In *The Twelfth International Conference on Learning Representations*, 2024.

Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. Unbiased learning-to-rank with biased feedback. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pp. 781–789, 2017.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 6769–6781, 2020.

Elahe Khatibi, Ziyu Wang, and Amir M. Rahmani. CDF-RAG: Causal dynamic feedback for adaptive retrieval-augmented generation. *arXiv preprint arXiv:2504.12560*, 2025.

Omar Khattab and Matei Zaharia. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 39–48, 2020.

Romain Laroche, Paul Trichelair, and Rémi Tachet des Combes. Safe policy improvement with baseline bootstrapping. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 3652–3661, 2019.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020.

Marc Lipsitch, Eric J. Tchetgen Tchetgen, and Ted Cohen. Negative controls: A tool for detecting confounding and bias in observational studies. *Epidemiology*, 21(3):383–388, 2010.

Peiyang Liu, Qiang Yan, Ziqiang Cui, Di Liang, Xi Wang, and Wei Ye. Beyond semantic relevance: Counterfactual risk minimization for robust retrieval-augmented generation. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1232–1243, 2026. doi: 10.1145/3805712.3809631.

James McInerney, Brian Brost, Praveen Chandar, Rishabh Mehrotra, and Benjamin Carterette. Counterfactual evaluation of slate recommendations with sequential reward interactions. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1779–1788, 2020. doi: 10.1145/3394486.3403229.

Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. GAIA: A benchmark for general AI assistants. In *The Twelfth International Conference on Learning Representations*, 2024.

Rodrigo Nogueira and Kyunghyun Cho. Passage re-ranking with BERT. *arXiv preprint arXiv:1901.04085*, 2019.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*, 2023. URL `https://arxiv.org/abs/2310.08560`.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 2023.

Huaiyu Qin, Chunyu Wei, Yueguo Chen, and Yunhai Wang. Counterfactual reasoning for retrieval-augmented generation. In *The Fourteenth International Conference on Learning Representations*, 2026.

James M. Robins, Miguel A. Hernán, and Babette Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, 2000. doi: 10.1097/00001648-200009000-00011.

Paul R. Rosenbaum. *Observational Studies*. Springer, 2 edition, 2002.

Ayush Sawarni, Jikai Jin, Justin Whitehouse, and Vasilis Syrgkanis. Policy learning with abstention. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, 2026.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik R. Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 8634–8652, 2023.

Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 814–823, 2015.

Eric J. Tchetgen Tchetgen, Isabel R. Fulcher, and Ilya Shpitser. Auto-g-computation of causal effects on a network. *Journal of the American Statistical Association*, 116(534):833–844, 2021. doi: 10.1080/01621459.2020.1811098.

Philip S. Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High-confidence off-policy evaluation. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 3000–3006, 2015a.

Philip S. Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High confidence policy improvement. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pp. 2380–2388, 2015b.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *Transactions on Machine Learning Research*, 2024.

Nengbo Wang, Xiaotian Han, Jagdip Singh, Jing Ma, and Vipin Chaudhary. CausalRAG: Integrating causal graphs into retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 22680–22693, 2025. doi: 10.18653/v1/2025.findings-acl.1165.

Zidi Xiong, Yuping Lin, Wenya Xie, Pengfei He, Zirui Liu, Jiliang Tang, Himabindu Lakkaraju, and Zhen Xiang. How memory management impacts LLM agents: An empirical study of experience-following behavior. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 623–645, 2026. doi: 10.18653/v1/2026.acl-long.27.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*, 2023.

Yi Yu, Liuyi Yao, Yuexiang Xie, Qingquan Tan, Jiaqi Feng, Yaliang Li, and Libing Wu. Agentic memory: Learning unified long-term and short-term memory management for large language model agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21457–21483, 2026. doi: 10.18653/v1/2026.acl-long.981.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. MemoryBank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19724–19731, 2024. doi: 10.1609/aaai.v38i17.29946.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. WebArena: A realistic web environment for building autonomous agents. In *The Twelfth International Conference on Learning Representations*, 2024.

A Randomization and Negative-Control Diagnostics

Let N_e be the fixed number of eligible opportunities in episode e. Under independent mask probability p=0.15 and a uniform max-3 cap, an opportunity's retention probability is

$$q_e = p \sum_{j=0}^{N_e-1} \binom{N_e-1}{j} p^j (1-p)^{N_e-1-j} \min\left(1, \frac{3}{j+1}\right).$$

With A_o=1 for forced include, A_o=0 for forced exclude, and arm probability $q_{e(o)}/2$, the leave-cluster-out label for development-fixed TraceGate-score bin g is

$$\hat{\tau}_g^{(-c)} = \frac{1}{|\mathcal{U}_g^{(-c)}|} \sum_{o \in \mathcal{U}_g^{(-c)}} \left[\frac{\mathbf{1}(A_o = 1) Y_o}{q_{e(o)}/2} - \frac{\mathbf{1}(A_o = 0) Y_o}{q_{e(o)}/2} \right].$$

Here $\mathcal{U}_g^{(-c)}$ is the complete eligible frame outside target task cluster c; unretained opportunities contribute zero to both arm totals.

Two-sample KS tests show close alignment between randomized and non-randomized decisions on the key pre-treatment features: tool reliability Z_t (KS=0.038, p=0.74), task difficulty (KS=0.062, p=0.47), and policy uncertainty (KS=0.054, p=0.53). Negative controls reinforce the calibration: among 200 known-irrelevant fragments, the false discovery rate for $|s(m)| > 1.0$ is 2.8% ± 0.7%, and placebo interventions yield ATE -0.11 ± 0.28 pp (Lipsitch et al., 2010). The remaining disagreements localize to weak overlap/positivity (38.4%), heavy-tailed verifier latency (22.1%), and misspecified Z_t (18.7%).

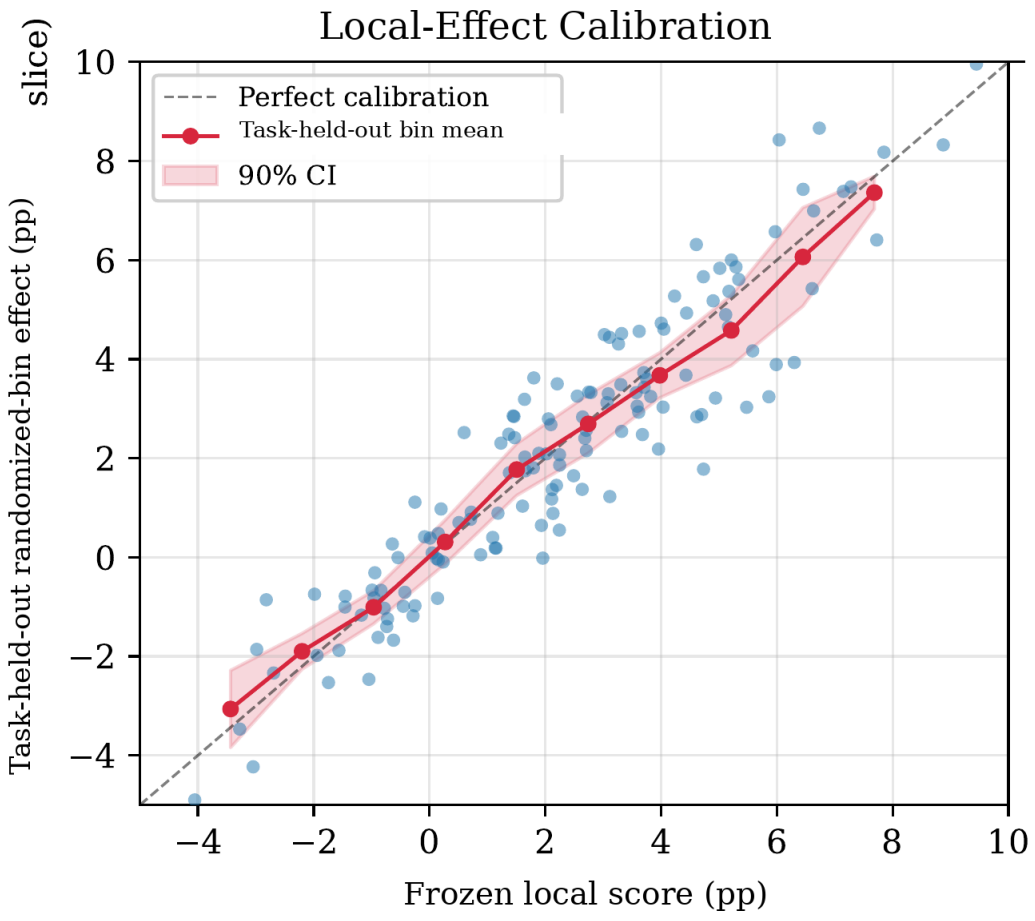


Figure 3: Frozen local scores versus inherited task-held-out Horvitz–Thompson bin effects. Blue marks are a fixed display sample of retained opportunities; repeated vertical values reflect their shared development-fixed bin label. Red markers and bands show bin means and 90% task-cluster intervals. All reported calibration metrics use the complete 15,600-opportunity frame.

Figure 3 visualizes the same score-to-held-out-bin comparison used by the sign-agreement and MAE definitions in Section 5; neither axis represents an individual pair of potential outcomes. The negative-control and placebo results support the label construction independently of the observational nuisance models.

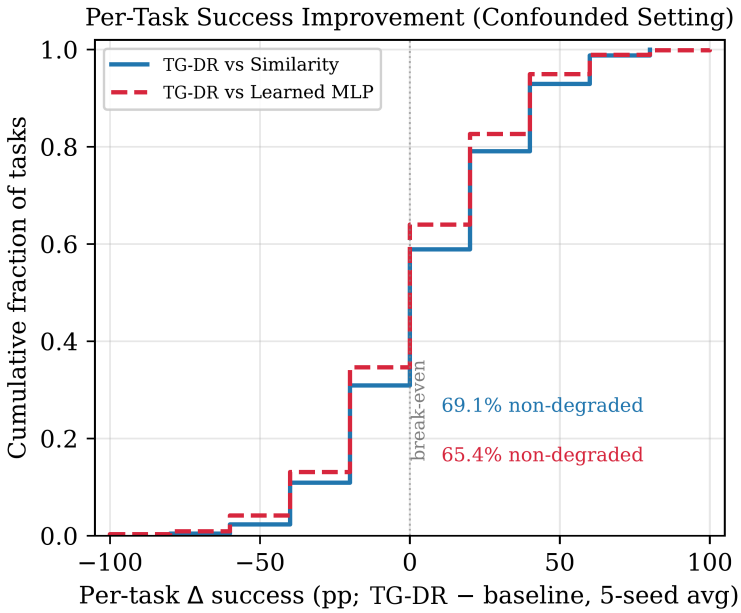


Figure 4: Per-task success-improvement CDF in the confounded setting. Five binary-seed outcomes quantize task deltas in 20 pp increments.

The staircase in Figure 4 follows from five binary outcomes per task. The reported 69.1% and 65.4% non-degradation frequencies are computed from the paired task table rather than read from interpolated curve locations.

B Extended Baseline Comparison

Table 4 reports every evaluated scorer under the same candidate pool and budgets. Neural rerankers already form a strong comparator band, while TG-DR improves all three settings and retrieves fewer fragments. Table 5 separates the confounded-only context count from latency pooled across settings.

Table 4: Full verifier-pass comparison across controlled settings.

Method	Confounded	Shift	Trap 25%
Similarity	37.8 ± 2.1	40.6 ± 1.9	35.7 ± 2.3
Similarity + tool state	39.4 ± 1.8	42.7 ± 1.7	37.6 ± 2.0
Stratified	40.1 ± 1.9	43.2 ± 1.8	38.0 ± 2.1
Cross-encoder	41.3 ± 1.5	44.5 ± 1.4	39.2 ± 1.7
Learned MLP	41.6 ± 1.4	44.8 ± 1.3	39.6 ± 1.5
LLM judge	42.4 ± 1.3	45.3 ± 1.2	40.1 ± 1.6
History router	42.1 ± 1.4	45.7 ± 1.3	40.8 ± 1.5
Randomized + IPS	39.2 ± 2.3	42.4 ± 2.1	38.3 ± 2.4
TG-DR	44.2 ± 1.6	47.9 ± 1.5	43.4 ± 1.7
TG-SLA	44.9 ± 1.7	48.6 ± 1.5	44.1 ± 1.8
TG-full	45.8 ± 1.9	49.7 ± 1.7	45.3 ± 1.9

C Matched-Success Analysis

We adjusted TG-DR's retrieval budget k to match similarity retrieval within 0.5 percentage points. In the confounded setting, TG-DR matches 37.8% success with 12.4 fragments per episode versus 21, a

Table 5: Efficiency scopes separated from success results. Fragment counts are confounded-setting only; p99 latency is pooled.

Method	Fragments/episode	Pooled p99 (s)
Similarity	21	1.4
LLM judge	17	2.8
History router	17	1.8
Same-pool CRM/DR	16	2.5
TG-DR	15	2.3
TG-SLA	15	5.6
TG-full	14	42.3

41.0% reduction. Under shift, it matches 40.6% with 11.8 versus 19 fragments, a 37.9% reduction. The arithmetic mean of these setting-level reductions is 39.5%.

D Ablation Studies

Table 6 extends the same-feature main-paper ablation with the probe-enabled reference. Propensity correction, the DR term, and logged intervention variables jointly produce the local-ordering gain; bounded probes add a separate accuracy/latency trade-off.

Table 6: Component ablations in the confounded setting. Gain is measured against similarity retrieval.

Variant	Gain (pp)	Spearman	Interpretation
TG-full	$+8.0 \pm 1.9$	$+0.29 \pm 0.04$	DR + propensity + probes
TG-DR (no probes)	$+6.4 \pm 1.6$	$+0.26 \pm 0.04$	failure boundary
w/o propensity model	$+3.6 \pm 1.3$	$+0.09 \pm 0.05$	biased under confounding
w/o DR correction	$+4.2 \pm 1.4$	$+0.13 \pm 0.04$	higher variance
withhold Z_t from nuisances	$+3.9 \pm 1.5$	$+0.11 \pm 0.05$	loses tool-confound control

E Set Interaction and Intervention-Logging Diagnostics

Interference diagnostics. Table 7 reports a model-based set-interaction diagnostic on the randomized slice. At k=16, the median residual is 0.41 pp and p95 is 1.32 pp; the 2.14 pp p99 tail activates the conservative gate. Growth at k=32 motivates the deployed set-size boundary, while direct randomized slates independently test the selected k=16 operating point.

Table 7: Model-based interaction residual by retrieval set size on the randomized audit slice.

k	Med. residual	p95 residual	p99 residual
8	0.28	0.94	1.61
16	0.41	1.32	2.14
32	0.63	1.78	3.02

Z_t misspecification. Table 8 re-estimates nuisance models with corrupted or withheld intervention values. The record-availability flag remains present in every row, so the 17.1% fallback rate reflects degraded propensity/ESS support rather than the mandatory missing-record branch. This differs from deleting Z_t from the decision record, which would trigger fallback for every affected candidate.

F Sensitivity Analysis

Table 9 measures robustness to hidden assignment bias. Gains retain a positive lower bound through $\Gamma = 1.5$; at $\Gamma = 2.0$, the widened interval reduces sign certainty and routes more online decisions

Table 8: Robustness to Z_t misspecification in the confounded setting.

Z_t condition	TG-DR (%)	Bin-sign agr.	Fallback
Full Z_t	44.2 ± 1.6	87.4%	5.9%
50% noise	43.1 ± 1.7	83.9%	12.3%
Withhold Z_t from nuisances	42.4 ± 1.8	81.6%	17.1%

away from promotion. Candidate-set size shows the expected support trade-off: top-64 yields $+5.1 \pm 1.3$ pp, top-128 yields $+6.4 \pm 1.6$ pp, and top-256 yields $+6.8 \pm 1.7$ pp with higher latency.

Table 9: Hidden-confounding sensitivity. FDR is the worst-case negative-control rate over the admissible propensity-odds set.

Sensitivity Γ	Lower bd. (pp)	Spearman	Worst-case FDR (%)	Admissible odds ratio
1.0	$+6.1 \pm 0.9$	$+0.26 \pm 0.04$	2.8	1
1.2	$+4.8 \pm 1.0$	$+0.21 \pm 0.04$	3.7	$[1/1.2, 1.2]$
1.5	$+3.1 \pm 1.1$	$+0.14 \pm 0.05$	5.2	$[1/1.5, 1.5]$
2.0	$+0.8 \pm 1.3$	$+0.06 \pm 0.06$	7.9	$[1/2, 2]$

G Implementation Details

Similarity generation uses FAISS HNSW and takes 14.2 ms p50 / 36.8 ms p95 per retrieval step; propensity plus outcome inference takes 4.1 ms / 10.3 ms, and DR scoring takes 5.7 ms / 14.6 ms. TG-DR disables probes, TG-SLA caps them, and TG-full exposes their 42.3 s pooled p99 tail. The outcome model is a two-layer 256-hidden-unit MLP trained by MSE; the propensity model is XGBoost with 128 depth-6 trees. Both train on the preceding session and freeze for the next window. The typed DAG stores pre-treatment observations, retrieval decisions, tool state, interventions, outcomes, diagnostics, and actions. Fragments are canonicalized from tool-error signatures, successful fixes, and invariant constraints, then deduplicated by normalized text plus a type tag.

Upstream assets and terms. The reported experiments use only the upstream assets named in Section 5. The model stack is Meta Llama 3.1 instruction models (70B for the main agent and 8B for the judge baseline) under the Llama 3.1 Community License and Acceptable Use Policy, `bge-large-en-v1.5` under MIT, and `bge-reranker-v2-m3` under Apache 2.0. The repository-repair collection uses SWE-bench-format patch verification; WebArena is under Apache 2.0, and GAIA follows its gated benchmark-access terms. The reported software stack is vLLM, FAISS, PyTorch, and XGBoost under their respective upstream licenses.

H Hyperparameter Sensitivity

Across the implementation sweeps in Table 10, gain versus similarity changes by at most 0.8 pp. Fallback responds more strongly, showing that support thresholds control coverage while preserving the central utility result. The hidden-bias sweep is reported separately in Table 9, where larger Γ widens the admissible propensity-odds set.

I Reporting Conventions and Comparison Scope

Units and terminology. An *opportunity* is one candidate at one pre-treatment retrieval state; a *retained edit* is an opportunity surviving the max-3 randomization cap; a *promoted edit* passes every TraceGate branch; an *episode* is one task–seed execution; a *task cluster* contains all seeds, episodes, and edits for one task; and a *slate* is the complete k-fragment retrieval action. Randomized-bin effects label opportunities only for calibration, while direct-slate effects estimate the deployed policy contrast. We report success as verifier-confirmed task pass rate averaged over five seeds. Differences in success rate, effects, and calibration error are in percentage points (pp); latency is end-to-end wall-clock time.

Table 10: Hyperparameter sensitivity in the confounded setting.

Parameter	Value	Gain (pp)	Bin-sign agr. (%)	Fallback (%)	Notes
$\pi_{\min}$	0.01	$+6.1 \pm 1.7$	86.8	4.2	looser clipping
	0.02 (default)	$+6.4 \pm 1.6$	87.4	5.9	
	0.05	$+6.2 \pm 1.5$	88.1	8.7	more conservative
Clip C	5	$+6.0 \pm 1.7$	87.1	6.3	more weight clipping
	10 (default)	$+6.4 \pm 1.6$	87.4	5.9	
	20	$+6.3 \pm 1.7$	87.2	5.6	less clipping
ESS threshold	15	$+6.6 \pm 1.8$	86.1	3.8	more disagreement
	25 (default)	$+6.4 \pm 1.6$	87.4	5.9	
	50	$+5.8 \pm 1.4$	88.7	11.4	more conservative

Detailed feature access. All methods receive pre-treatment information only and share the task instruction, compact state summary, recent tool history, candidate pool, and harness. Similarity and the cross-encoder consume text; Learned MLP and History router consume structured state; the LLM judge consumes the textualized state and compact history; the same-pool CRM/DR learner uses the complete pre-treatment surface and logged exposure. TG-DR combines S_t, candidate features, $\hat{Q}$, $\hat{\pi}$, and the action gates. Every method uses the same retrieval query: task instruction, latest tool-error signature when present, and a state summary capped at 256 tokens.

J Neural Comparator Specifications

The LLM judge and History router most often define the strongest non-causal comparator in Table 3. They provide stringent semantic and trajectory-conditioned controls alongside the same-pool logged-feedback learner.

LLM-judge reranker. We use Llama-3.1-8B-Instruct served via vLLM 0.5.4 (FP16, tensor-parallel 1, A100 80 GB) as a relevance judge. For each candidate fragment m in the top-128 similarity set, we construct a prompt: "`Given the current task state and tool history, rate the relevance of the following memory fragment to the current retrieval need on a scale of 1--10. Fragment: [m]. Rating:`". We extract the integer rating and rank by it. Batched inference achieves 45 fragments/s at FP16 precision; its pooled end-to-end episode latency has a 2.8 s p99. The LLM-judge captures semantic nuances missed by embedding cosine (e.g., relevance conditioned on error patterns), but it cannot distinguish correlation from causation because it scores each fragment independently without modeling how tool reliability confounds outcomes.

History-aware router. We train a 4-layer Transformer encoder (hidden dim 256, 4 heads, dropout 0.1) on sequences of the last 8 (state, action, tool-result, retrieval-outcome) tuples. The model predicts binary success for each candidate fragment via a linear head on the [CLS] token embedding concatenated with the candidate's bge-large-en-v1.5 vector. Training uses preceding-session logs (same as Learned MLP) with binary cross-entropy loss and early stopping on a held-out 20% split. The history-aware router conditions on recent trajectory context, which partially captures policy shift; however, it does not model intervention variables or apply DR corrections, so it conflates tool-state correlation with causal effect.

K Coverage–Disagreement Operating Points

All rows use the same 15,600 retained randomized occurrences. Gate ablations keep the remaining TG-DR thresholds fixed; ordinary-scorer cutoffs are fixed on development data to the TG-DR promoted-coverage neighborhood. "Bin disagreement" means disagreement with the task-held-out TraceGate-score-bin effect defined in Section 5. The table reports descriptive point estimates; no pairwise interval is claimed for cross-method differences.

Table 11: Selected coverage and disagreement with task-held-out TraceGate-score-bin signs at frozen operating points. Ordinary scorers are coverage-matched to TG-DR; cross-method rows are descriptive common-reference diagnostics.

Method	Selected	Coverage	Bin disagreement	Rate
TG-DR	9,978	64.0%	604	6.1%
w/o overlap gate	10,998	70.5%	957	8.7%
w/o sensitivity gate	10,717	68.7%	793	7.4%
w/o uncertainty gate	11,263	72.2%	1,104	9.8%
Score-only DR	9,999	64.1%	1,090	10.9%
Same-pool CRM/DR	10,015	64.2%	921	9.2%
Learned MLP	9,986	64.0%	1,378	13.8%
Similarity + state	10,003	64.1%	1,651	16.5%

Table 12: Utility and routing consequence for coverage–disagreement operating points. Verifier gain is the equal-setting average against the setting-wise strongest non-causal comparator; probe rate is zero throughout.

Method	Verifier gain (pp)	Fallback rate
TG-DR	+2.2	7.1%
w/o overlap gate	+1.8	4.0%
w/o sensitivity gate	+1.9	6.3%
w/o uncertainty gate	+1.5	3.8%
Score-only DR	+1.3	0.0%
Same-pool CRM/DR	+1.5	0.0%
Learned MLP	+0.8	0.0%
Similarity + state	+0.5	0.0%

Overlap removal admits unsupported cases; uncertainty removal admits the largest number of ambiguous edits; sensitivity removal produces a smaller but distinct disagreement increase. The full rule occupies the favorable low-disagreement region rather than tracing a mechanically monotone curve.

L Cross-Family Replication

The Qwen2.5-72B-Instruct replication evaluates 650 tasks and 3,250 episodes per setting. Five paired seeds, development-fixed thresholds, calibration bins, the fragment pool, top-128 generator, prompt templates, tools, and verifier budgets match the Llama-3.1 evaluation. Model-dependent nuisance functions and learned comparator policies, including same-pool CRM/DR parameters, are refit on Qwen preceding-window logs; their development-selected hyperparameters remain frozen.

Table 13: Qwen2.5 verifier-pass rate (mean$\pm$seed standard deviation).

Method	Confounded	Shift	Trap 25%
Best non-causal	41.8 ± 1.5	45.0 ± 1.4	39.7 ± 1.6
Same-pool CRM/DR	42.9 ± 1.5	46.1 ± 1.4	41.2 ± 1.7
Score-only DR	42.7 ± 1.6	45.9 ± 1.5	41.0 ± 1.7
TG-DR	43.8 ± 1.6	47.0 ± 1.5	42.5 ± 1.8

Under the common TraceGate-score-bin diagnostic, TG-DR has the highest agreement point estimate; its verifier success is also highest in every setting while context and latency remain comparable. The common-bin comparison is descriptive because the strata are not separately rebuilt for each method. Relative to the same-pool learner, task win/tie/loss counts are 204/361/85, 198/370/82, and 221/342/87; fallback rates are 6.6%, 10.1%, and 7.8%.

Table 14: Qwen2.5 cross-family diagnostic and efficiency metrics. The TG-DR MAE "±" value is a 95% task-cluster bootstrap half-width; fragment counts are confounded-setting only and latency is pooled.

Method	Common-bin agr.	MAE (pp)	Fragments	p99 (s)
Best non-causal	—	—	17	2.9
Same-pool CRM/DR	83.7%	1.39	16	2.6
Score-only DR	82.4%	1.42	15	2.4
TG-DR	85.9%	1.27 ± 0.16	15	2.5

M Failure-Boundary Evidence Matrix

The two aligned panels connect each protected decision to its failure capability, assumptions, comparator, operational effect, and manuscript evidence.

Table 15: Evidence matrix, panel A: protected object and trust boundary.

Failure surface	Protected object	Capability	Trust assumption	Property / comparator
Misspecified tool state	local edit	corrupted reliability mode	complete pre-treatment Z_t	fallback / nuisance ablation
Weak overlap	local edit	rare include/exclude arm	ESS$\geq$25, $\hat{\pi}_t \geq 0.03$	support / similarity+state
Hidden confounding	effect sign	latent odds shift	fixed $\Gamma^\star$=1.5 stress	interval stability / CRM-DR
Set interference	k-fragment action	interacting fragments	residual inside size gate	model diagnostic / direct slate
Policy shift	retrieval policy	changed constraints	frozen-window stationarity	fixed fallback / History router
Correlational trap	selected fragment	lexical decoy	logged intervention state	randomized-bin agreement / cross-encoder
Probe latency	episode completion	heavy verifier tail	explicit rate/time caps	SLA / no-probe TG-DR
Blocked transfer	held-out task	task/repository exposure	disjoint folds	transfer / CRM-DR

The matrix maps each tested failure capability to its gate, assumption, measured consequence, and paper-internal evidence trace.

N Same-Pool Logged-Feedback Comparator

The same-pool CRM/DR learner uses the complete TraceGate pre-treatment feature vector and a two-layer hidden-256 policy scorer $f_\theta(S_t, \phi(m))$. One training sample t is one logged candidate opportunity m_t with binary inclusion $I_t \in \{0, 1\}$, episode outcome Y_t, and logged marginal inclusion propensity $\hat{\pi}_t$. Its stochastic training target is the Bernoulli candidate-inclusion marginal $p_{\theta,t} = \sigma(f_\theta(S_t, \phi(m_t)))$; it is not a normalized categorical distribution or a slate likelihood. Deterministic inference takes the 16 highest logits from the shared top-128 pool, so training and inference use the

Table 16: Evidence matrix, panel B: stress setting, cost, observed failure, and evidence trace.

Failure surface	Adaptive / held-out setting	Utility or overhead	Observed failure	Evidence
Misspecified tool state	50% noise and nuisance-model withholding	fallback rises 5.9% to 17.1%	bin-sign agreement falls to 81.6%	Table 8
Weak overlap	randomized top-8	64.0% promoted coverage	38.4% of disagreements	Fig. 2
Hidden confounding	$\Gamma = 1.0$–2.0	lower bound +6.1 to +0.8 pp	worst-case negative-control FDR rises	Table 9
Set interference	$k = 8/16/32$	model-based median 0.28/0.41/0.63 pp	p99 reaches 3.02 pp	Table 7
Policy shift	frozen policy constraints	+2.2 pp vs. History router	9.4% overlap fallback	Table 3
Correlational trap	25% trap injection	+2.6 pp vs. History router	misleading lexical score	Table 3
Probe latency	3% and 25% probes	p99 5.6 and 42.3 s	heavy-tailed verifier	Table 5
Blocked transfer	650 task / 54 repo folds	positive point estimates in all settings	shift interval crosses zero	Table 18

same candidate scores but the training objective remains candidate-marginal. On each preceding window it maximizes

$$\widehat{V}_{\mathrm{CRM/DR}}(\theta) = \frac{1}{n}\sum_{t=1}^{n} \psi_t(\theta) - \lambda\sqrt{\widehat{\mathrm{Var}}(\psi_t)/n} - \eta\|\theta\|_2^2.$$

Writing $\hat{Q}_t(a) = \hat{Q}(S_t, \phi(m_t), a)$ and $\hat{\pi}_t = \hat{\pi}_{m_t}(S_t)$, its DR term is

$$\begin{aligned}\psi_t(\theta) = {} & p_{\theta,t}\hat{Q}_t(1) + (1 - p_{\theta,t})\hat{Q}_t(0) \\ & + \left[\frac{p_{\theta,t} I_t}{\max(\hat{\pi}_t, 0.02)} + \frac{(1 - p_{\theta,t})(1 - I_t)}{\max(1 - \hat{\pi}_t, 0.02)}\right](Y_t - \hat{Q}_t(I_t)).\end{aligned}$$

λ, η, and early stopping are selected on the development partition, then frozen; nuisance models and policy parameters are refit only on the preceding 500-episode window. Both inverse-propensity denominators are clipped below at 0.02 in the displayed DR term. The learner has no overlap, sensitivity, uncertainty, or probe branch. A common input-validity rule delegates to similarity+tool-state only when a required pre-treatment feature is absent; this availability fallback does not inspect ESS, intervals, Γ, or score uncertainty. For the local frontier, whose required features are complete, the learner's promoted event is $f_\theta(m) \geq c_{\mathrm{CRM}}$, with c_{CRM} fixed on development data to the 64% coverage neighborhood. Table 17 therefore compares a candidate-marginal logged-feedback learner on the same pool, state, windows, and budgets; it is not presented as a slate-level DR estimator.

Table 17: Same-pool CRM/DR comparator with matched state, pool, and budgets.

Method	Confounded	Shift	Trap 25%
Best non-causal	42.4 ± 1.3	45.7 ± 1.3	40.8 ± 1.5
Same-pool CRM/DR	43.5 ± 1.6	47.0 ± 1.5	42.3 ± 1.7
TG-DR	44.2 ± 1.6	47.9 ± 1.5	43.4 ± 1.7

Under the common TraceGate-score-bin diagnostic, the same-pool learner reaches 84.6% agreement and 1.33 pp MAE, versus 87.4% and 1.18 pp for TG-DR; this is a descriptive common-reference comparison. It retrieves 16 fragments per confounded-setting episode and reaches pooled p99 2.5 s.

Paired permutation tests give 0.041, 0.083, and 0.028; the confounded and trap effects resolve, while the shift point estimate remains positive with greater uncertainty.

O Blocked Task and Repository Transfer

The blocked evaluation aggregates disjoint held-out folds so each of the 650 tasks contributes exactly once. Repository-repair folds group all 54 repositories without cross-fold exposure. Thresholds and comparator identities are selected on development data before fold evaluation; the strongest non-causal baseline is fixed per setting (LLM judge for confounded, History router for shift and trap), and nuisance models and the same-pool learner train without the evaluated task or repository.

Table 18: Blocked task/repository validation with frozen thresholds. Comparator identities are fixed per setting before scoring.

Method	Confounded	Shift	Trap 25%
Best non-causal	42.2 ± 1.5	45.4 ± 1.4	40.9 ± 1.7
Same-pool CRM/DR	42.4 ± 1.6	45.8 ± 1.5	41.2 ± 1.8
TG-DR	43.0 ± 1.7	46.3 ± 1.6	41.9 ± 1.8

Across 650 held-out tasks and 54 repository groups, TG-DR reaches 82.9% task-held-out TraceGate-bin agreement and 1.41 pp MAE. Under the same descriptive common-bin diagnostic, the same-pool learner reaches 81.5% and 1.49 pp; its input-availability fallback rates are 7.6%/10.2%/9.1%, not TraceGate-style support deferrals. TG-DR minus same-pool paired intervals are [0.1, 1.2], [-0.1, 1.1], and [0.1, 1.4], with permutation $p = 0.038$, 0.112, and 0.031. TG-DR fallback rises to 8.1%, 10.7%, and 9.8%; confounded and trap effects exclude zero, while the shift estimate remains positive with an interval crossing zero.

P Direct Set-Level Intervention at k=16

For each setting, the direct experiment samples 960 frozen pre-treatment retrieval states from the same 650-task population before any method scores the shared candidate pool. Each state is assigned with prespecified probability $1/3$ to TG-DR, the development-fixed strongest non-causal method, or the same-pool CRM/DR learner. The complete $k{=}16$ slate is the randomization unit: an assigned method constructs all positions, the agent resumes from the saved state, and the episode verifier supplies the binary outcome. Repeated states from one task remain in the same inference cluster. Table 19 reports Horvitz–Thompson arm-mean point estimates under the known assignment probabilities; policy-level interval claims use the larger paired episode analysis in Table 3.

Table 19: Descriptive direct $k{=}16$ effects in pp from 960 total three-arm randomized slates per setting.

Setting	vs. best non-causal	vs. same-pool	Total slates
Confounded	+1.6	+0.7	960
Policy shift	+1.9	+0.5	960
Trap 25%	+2.1	+0.8	960

The direct point estimates align with the paired episode effects in all three settings. As a separate model check, fitted slate-response contrasts agree in sign with summed local predictions 88.7% of the time against the strongest non-causal comparator and 84.9% against the same-pool learner; their absolute setting-level gaps are 0.4, 0.5, and 0.6 pp.