TRUCE: CONFLICT-CONDITIONED INTERVENTIONAL ROUTING FOR TOOL AGENTS AFTER UPDATES

Anonymous authors
Paper under double-blind review

ABSTRACT

After a tool or fact update, an agent can simultaneously retrieve a current semantic record, an obsolete procedural recipe, and a successful episodic trace for the retired behavior; exposing all three can turn relevance into the wrong next call. Aggregate retrieval accuracy does not answer the resulting systems question: when is the conditional reliability gain from resolving typed disagreement large enough to justify randomized estimation, uncertainty calibration, and routing latency? The *conflict–value profile* answers this question by indexing held-out policy gain by pre-routing disagreement, support, latency, and exploration cost. Its realization, TRUCE (Typed-memory Routing Under Conflicting Evidence), estimates same-background singleton and pairwise store contrasts from exact post-clipping action propensities, gates weakly supported scores, composes a latency-feasible subset, and resolves revision contradictions. Across three public update benchmarks, TRUCE gains $+2.7/+2.8/+2.7$ success points over an exposure-matched supervised router in conflict-heavy episodes, while low-conflict differences are $+0.3/+0.4/+0.2$ points; gains over the strongest same-feature subset policy are $+2.7/+2.3/+2.3$. Episode-wise randomization of store identities reduces the learned-router advantage to $+0.12$ points (95% CI $[-0.41, +0.63]$), and greedy routing agrees with exact fitted-score search on 94.1%/93.4% of public-core decisions while finishing within 0.2/0.2 success points. At 52% of the reference exploration-and-training cost, the full stack retains 85% of its public-core lift, making typed conflict a measurable trigger for spending routing complexity after updates.

1 INTRODUCTION

An updated tool can leave an agent with three plausible histories of the same action. The semantic store records the new schema, the procedural store preserves yesterday's call recipe, and the episodic store contains a successful execution under the retired signature. Because these sources play different roles (Park et al., 2023; Mialon et al., 2023; Zhong et al., 2024), a revision can turn useful memory separation into an action-level conflict: relevance alone cannot determine whether combining the sources will corroborate the update or revive an obsolete call.

The effect is measurable before choosing a route. Across WIKIREV-PUBLIC, EDITQA-CONFLICT, and API-family-disjoint APIMIGRATE-PUBLIC, the advantage of TRUCE over an exposure-matched router is 2.7–2.8 success points in conflict-heavy episodes but 0.2–0.4 points in low conflict. A separate ten-bin diagnostic rises from $+0.3$ to $+4.1$ points ($\rho = 0.94$ over bin means); when store identities are randomized episode by episode while content is preserved, the learned-router gap falls to $+0.12$ points. The benchmark-wise boundary, graded response, and identity intervention make typed disagreement—rather than a uniform accuracy offset—the discriminating variable.

Current retrieval and tool-agent systems optimize evidence access or aggregate task performance (Lewis et al., 2020; Izacard & Grave, 2021; Asai et al., 2024; Schick et al., 2023; Qin et al., 2024; Yao et al., 2023). They do not expose the conditional payoff of intervention after independently useful stores disagree. That payoff matters because specialized control consumes randomized decisions for estimation, inference latency for subset selection, and capacity for uncertainty calibration. We formalize it as the *conflict–value profile*: held-out policy gain indexed by pre-routing conflict under a fixed latency cap and an explicitly measured exploration-and-training budget.

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.

TRUCE realizes the profile through one identification-to-action path. A randomized wrapper records exact probabilities for latency-clipped store subsets. Cross-fitted action-level AIPW scores then yield same-background singleton and pairwise contrasts; episode-clustered uncertainty excludes unsupported effects; greedy composition enforces the latency cap; and a recency-aware corroboration veto resolves contradictions among selected items. Exact fitted-score search and an eight-action same-log DR learner test the deployed policy layer rather than only the local scores.

The paper contributes three linked technical objects:

- *Conflict–value profiles* connect pre-routing typed-store disagreement to conditional reliability gain, support, latency, and exploration cost.
- TRUCE identifies post-clipping same-background store effects and composes them through calibrated uncertainty gates, a latency-feasible route, and revision-aware corroboration.
- A falsification-oriented evaluation links continuous conflict response, matched policy controls, type-label placebo, component attribution, exact fitted-score optimization, and cost saturation across factual and API updates.

2 Related Work

Retrieval and tool agents. RAG, FiD, RETRO, ColBERT, Self-RAG, and MemoRAG improve evidence access through retrieval, interaction, reflection, or long-context memory (Lewis et al., 2020; Izacard & Grave, 2021; Borgeaud et al., 2022; Khattab & Zaharia, 2020; Asai et al., 2024; Qian et al., 2025); Toolformer, ToolLLM, ReAct, Gorilla, HuggingGPT, WebGPT, and WebArena extend tool selection and interactive evaluation (Schick et al., 2023; Qin et al., 2024; Yao et al., 2023; Patil et al., 2024; Shen et al., 2023; Nakano et al., 2021; Zhou et al., 2024). TRUCE instead measures the conditional value of changing a route after source disagreement is observable.

Typed and structured memory routing. ENGRAM separates episodic, semantic, and procedural stores; TAG mixes memory granularities; AMA combines retrieval, consistency checks, and refresh; MAGMA routes over relational views (Patel & Patel, 2025; Qing et al., 2026; Huang et al., 2026; Jiang et al., 2026). Their routing is relevance- or intent-conditioned. TRUCE's update-time object is different: supported action value after populated stores conflict, estimated under post-clipping subset randomization and exposed as a value–cost profile.

Memory updates. Model editing changes parametric associations, while continual learning preserves competence across updates (De Cao et al., 2021; Mitchell et al., 2022; Meng et al., 2022; 2023; Kirkpatrick et al., 2017; Lopez-Paz & Ranzato, 2017; Parisi et al., 2019). Here actionable versions coexist outside a fixed backbone, making source selection for the next call—not weight editing or replay—the control surface.

Logged subset decisions. Contextual bandits, IPS, DR, and DML motivate LinUCB-Ctx, SubsetPolicy-DML, and SameLog-DR controls with the same context or log (Auer et al., 2002; Li et al., 2010; Dudík et al., 2011; Jiang & Li, 2016; Su et al., 2020; Chernozhukov et al., 2018). TRUCE estimates background-matched store interactions before feasible composition; exact enumeration measures the deployment cost of that factorization.

3 Problem Formulation

Setup and decision outcome. At decision t in episode e, the agent observes pre-routing context $X_{et} = x$ and exposes a subset $A_{et} = a$ of typed stores $\mathcal{S} = \{s_1, \ldots, s_K\}$. The action space is the latency-feasible part of $2^{\mathcal{S}}$, including $\emptyset$. $Y_{et}(a; \rho)$ is successful completion of the current subgoal within $W = 5$ steps after focal exposure a, with later routes generated by the frozen continuation ρ. It measures the focal route's total effect; all decisions from an episode share one inference cluster.

Conflict–value profile. Let $c(X)$ be a pre-routing disagreement score, π a store-selection policy with latency cap Λ, and π_B^0 a comparator trained with the same exploration budget B. For conflict interval Q, define the held-out policy contrast

$$\mathcal{V}_\pi(Q; \Lambda, B) = \mathbb{E}\big[Y(\pi(X; \Lambda)) - Y(\pi_B^0(X; \Lambda)) \mid c(X) \in Q\big] . \tag{1}$$

In Equation (1), $Y(\cdot)$ abbreviates $Y(\cdot;\rho)$. The collection of $\mathcal{V}_\pi$ across conflict intervals, together with measured exploration-plus-training cost $K(B)$, is the conflict–value profile. Here $K(B)$ is A100-GPU time for randomized data collection plus nuisance and score fitting, normalized to the $B = 0.15$ full stack; route latency is reported separately. Profile points come from paired held-out episodes. The local estimands below supply TRUCE's routing scores.

Post-clipping action law. An exploration mark $E \sim \mathrm{Bernoulli}(B)$ triggers independent store proposals with probability $1/2$; infeasible proposals are clipped by uniformly removing stores until the validation-measured additive cost is at most Λ. The exact $q(a \mid x, E{=}1)$ sums every proposal-and-removal path ending at final action a (Appendix N). Only marked decisions enter AIPW fitting, so B controls sample size rather than the conditional propensity denominator; reported route latency is measured end to end. Write $m_a(x) = \mathbb{E}[Y(a;\rho) \mid X = x]$. For store s, let $\mathcal{G}_s(x)$ contain backgrounds b for which b and $b \cup \{s\}$ are feasible and supported, with fixed logger-induced law $g_s(b \mid x)$. Its same-background value is

$$\tau_s(x) = \sum_{b \in \mathcal{G}_s(x)} g_s(b \mid x)\{m_{b\cup\{s\}}(x) - m_b(x)\}. \tag{2}$$

For pair (s, s'), $\mathcal{G}_{ss'}(x)$ requires support for all four corresponding actions and uses the analogous fixed background law:

$$\begin{aligned}\phi_{s,s'}(x) = \sum_{b \in \mathcal{G}_{ss'}(x)} g_{ss'}(b \mid x)\big[&m_{b\cup\{s,s'\}}(x) - m_{b\cup\{s\}}(x) \\ &- m_{b\cup\{s'\}}(x) + m_b(x)\big].\end{aligned} \tag{3}$$

This definition compares additions to the same realized background after latency clipping; it does not replace the joint assignment by a marginal store indicator. Pair terms capture jointly harmful stores, while corroboration handles observed higher-order contradictions.

Identification assumptions.

(i) Post-clipping overlap: every action cell entering a contrast has $q(a \mid x, E{=}1) > 0$; the retained audit additionally requires marginal propensities in $[0.10, 0.90]$ and every required final-action propensity to be at least 0.01.

(ii) Consistency and continuation stability: the observed outcome equals $Y_{et}(A_{et};\rho)$, the same history-dependent ρ governs later routing under every focal action, and the subgoal outcome closes at completion or step W. There is no interference across episodes.

(iii) Sequential randomization: conditional on pre-routing X_{et}, the focal proposal and latency-clipping variables are independent of $\{Y_{et}(a;\rho)\}_a$. Dependence created by overlapping windows inside an episode is retained for cross-fitting and inference.

Budget constraints. Each routing decision incurs retrieval latency. We budget retrieval latency at Λ per decision and exploration at B per episode, where the latter is the fraction of decisions receiving randomized exposure.

Proposition 1 (post-clipping local effects; proof in Appendix N). Under assumptions (i)–(iii), finite second moments, and episode-grouped cross-fitting, the action-level AIPW construction below identifies $\tau_s = \mathbb{E}[\tau_s(X)]$ and $\phi_{s,s'} = \mathbb{E}[\phi_{s,s'}(X)]$ on paired support. With the known final-action probabilities and standard nuisance-rate conditions, their episode-clustered averages are asymptotically normal. Context-dependent routing scores are regressions of out-of-fold contrast pseudo-outcomes; Proposition 1 concerns the averaged local contrasts, while the deployed score is calibrated and evaluated as a policy.

4 TRUCE: Realizing the Conflict–Value Profile

TRUCE is organized around three requirements of the profile: matched post-clipping comparisons, conflict-sensitive composition, and a budget-feasible next action. Randomization creates supported subset contrasts; singleton and pair terms represent source-specific value and incompatible combinations; calibrated confidence gates retain supported effects; greedy selection respects latency; and corroboration uses revision order to resolve selected evidence.

4.1 Data Collection via Randomized Exposure

During data collection, the logger records the exploration mark, proposal, ordered removals, post-clipping action, conditional $q(a \mid x, E{=}1)$, context, and outcome. Only marked decisions enter nuisance and contrast fitting. A dataset-level marginal exposure floor of 0.10 prevents gross imbalance, while deployment checks every final-action cell required by a local contrast. In an unsupported context, TRUCE records an abstention and executes Unified Retriever, the revision-aware standard policy used throughout the study.

4.2 Doubly-Robust Store-Value Estimation

For every supported final subset a, let $\mu_a(x) = \mathbb{E}[Y \mid X = x, A = a, E = 1]$. The recorded post-clipping probability $q(a \mid x, E{=}1)$ and a cross-fitted outcome model give the action-level AIPW pseudo-outcome (Dudík et al., 2011)

$$\hat{\psi}_a = \hat{\mu}_a(x) + \frac{\mathbf{1}[A = a]}{q(a \mid x, E{=}1)}\{Y - \hat{\mu}_a(x)\}. \tag{4}$$

We construct singleton pseudo-outcomes as $\hat{\psi}_s = \sum_b g_s(b \mid x)(\hat{\psi}_{b\cup\{s\}} - \hat{\psi}_b)$ and pair pseudo-outcomes as

$$\hat{\psi}_{s,s'} = \sum_b g_{ss'}(b \mid x)(\hat{\psi}_{b\cup\{s,s'\}} - \hat{\psi}_{b\cup\{s\}} - \hat{\psi}_{b\cup\{s'\}} + \hat{\psi}_b).$$

Regressing these out-of-fold quantities on context yields $\hat{\tau}_s(x)$ and $\hat{\phi}_{s,s'}(x)$. All folds group entire episodes, so decisions from one trajectory never enter both nuisance training and pseudo-outcome evaluation (Chernozhukov et al., 2018). This joint-action construction preserves the common background in Equations (2)–(3) despite latency clipping.

4.3 Routing Policy

The *surrogate uncertainty-gated additive-plus-pairwise score* below uses store-value estimates $\hat{\tau}_s(x)$, whose uncertainties are $\hat{\sigma}_s(x)$, together with pairwise interaction estimates $\hat{\phi}_{s,s'}(x)$, whose uncertainties are $\hat{\sigma}_{s,s'}(x)$:

$$\begin{aligned} F(a; x) = &\sum_{s\in a} \hat{\tau}_s(x) \cdot \mathbf{1}[\hat{\tau}_s > \gamma\hat{\sigma}_s] \\ &+ \sum_{\{s,s'\}\subseteq a} \hat{\phi}_{s,s'}(x) \\ &\cdot \mathbf{1}[|\hat{\phi}_{s,s'}| > \gamma\hat{\sigma}_{s,s'}], \end{aligned} \tag{5}$$

subject to $\mathrm{lat}(a) \le \Lambda$. Each $\hat{\sigma}(x)$ is the SD across 500 episode-cluster bootstrap refits, scaled per singleton/pair family by training-fold pseudo-outcome residuals; Table 14 audits prediction coverage. Validation selects $\gamma = 1.0$ from $\{0, 0.5, 1.0, 1.5, 2.0\}$ under support and 200 ms constraints, with larger-γ tie breaking. Test episodes tune neither scale nor gate.

F combines the identified local inputs into a second-order policy surrogate, and held-out success evaluates the induced route. Greedy and Exact Score Search optimize the same F, making their 93.4–94.1% agreement a policy-realization test.

Deterministic corroboration veto. After at least two stores are selected, TRUCE compares their top items using entailment, contradiction, and revision recency. It removes an item contradicted by a fresher selected-store item when no selected partner agrees; if every item is removed, it returns the store with the largest gated singleton score. Public decomposition experiments isolate this step from DR-guided selection.

4.4 Training Details

Three-layer 512-unit MLP routers use a frozen Llama-3.1-8B context encoder, five recent tool outputs, task phase, store staleness, and pair contradictions; the tool agent uses the Llama-3.1-70B backbone described below. Adam trains the routers for 50 epochs at $3{\times}10^{-4}$ with batch size 256; cross-fitting uses five episode-grouped folds. Exact Score Search enumerates pre-veto $F(a; x)$; greedy and exact routes then receive identical corroboration for outcome and latency evaluation.

Algorithm 1 TRUCE Routing at Inference (additive+pairwise DR)

Require: Context x, stores $\mathcal{S}$, singleton value $\hat{\tau}$, pairwise interaction $\hat{\phi}$, uncertainties $\hat{\sigma}_{\cdot}, \hat{\sigma}_{\cdot,\cdot}$, threshold γ, latency budget Λ

if any required paired action cell is unsupported **then**
  return Unified Retriever
end if
Compute $\hat{\tau}_s(x), \hat{\sigma}_s(x)$ for each $s \in \mathcal{S}$ and $\hat{\phi}_{s,s'}(x), \hat{\sigma}_{s,s'}(x)$ for each pair
$\mathcal{C} \leftarrow \{s : \hat{\tau}_s(x) > \gamma \hat{\sigma}_s(x)\}$
Sort $\mathcal{C}$ by $\hat{\tau}_s(x)$ descending; initialize $a \leftarrow \emptyset$, $\ell \leftarrow 0$
for $s \in \mathcal{C}$ **do**
  $\Delta_s \leftarrow \hat{\tau}_s(x) + \sum_{s' \in a} \hat{\phi}_{s,s'}(x) \cdot \mathbf{1}[|\hat{\phi}_{s,s'}| > \gamma \hat{\sigma}_{s,s'}]$
  if $\Delta_s > 0$ **and** $\ell + \text{lat}(s) \leq \Lambda$ **then**
    $a \leftarrow a \cup \{s\}$; $\ell \leftarrow \ell + \text{lat}(s)$
  end if
end for
if $|a| \geq 2$ **then**
  Run deterministic corroboration veto on the top retrieved items
end if
return a

5 Experimental Setup

5.1 Pre-Routing Conflict and Outcomes

An outcome-blind parser maps the request and each top item to an entity–property or API–endpoint–field key. Equal-key payloads are labeled agreement, contradiction, or neutral under frozen value/schema rules; the newest relevant revision is selected only among retrieved records matching the request key, without target or post-route information. Appendix B specifies normalization, incompatibility, missingness, and ties.

Every benchmark then uses the outcome-blind score

$$c(x) = 0.6D(x) + 0.25S(x) + 0.15R(x).$$

D indicates any contradictory pair; S is the fraction of top items predating the newest match; and $R = \min\{1, \Delta_t / T_b\}$ normalizes their timestamp span by the training 95th percentile. Missing timestamps count toward S but not Δ_t; absent or neutral pairs set $D = R = 0$. Weights, normalizers, and cutpoints are training-frozen before router fitting. High/low thresholds are $0.70/0.18$ (WikiRev), $0.74/0.19$ (EditQA), and $0.71/0.18$ (ToolRouting); APIMigrate uses training-frozen balanced-cohort cutpoints. Feature acquisition counts toward latency; success requires the normalized revised fact or an executable post-migration call.

5.2 Controlled Mechanism Benchmark

Three source datasets supply the episodes used to construct ToolRouting-Bench: 1,420 episodes from API-Bank (Li et al., 2023), 1,598 episodes from ToolLLM (Qin et al., 2024), and 1,200 episodes from a custom multi-session assistant benchmark. We augment every episode with three typed memory stores: timestamped episodic tool traces, semantic facts carrying revision history and confidence, and procedural skill templates that can be reused.

We inject controlled staleness (20% of semantic entries), revision conflict (15% of episodes), and store-specific poisoning (5% per store). Corpus quartiles contain n=1,054/1,055 conflict-heavy/low-conflict episodes. The 12,000 routing decisions come exclusively from 2,953 training episodes. Model selection uses 632 validation episodes, and Table 3 uses the disjoint 633-episode test split: 158 conflict-heavy and 159 low-conflict episodes. No evaluation episode contributes an exploration decision.

5.3 EXTERNAL BENCHMARKS

Three fully public benchmarks evaluate the profile across two update families: two factual-revision tasks and one natural API-update task. Controlled TOOLROUTING-BENCH isolates the mechanism; complementary stress sets test additional store counts, natural traces, robustness, and the typed-label placebo.

- WIKIREV-PUBLIC (2,120 episodes, *fully public*): built from $86{,}412$ Wikidata revision diffs committed 2025-07-01 to 2026-01-31. Typed stores are paired in each episode with a revision-resolution QA target. A live contradiction between stores occurs in 28.1% of episodes; the slices use thresholds of $c \geq 0.70$ (n=530) for high conflict and $c \leq 0.18$ (n=530) for low conflict. The training log contains $7{,}200$ decisions.
- EDITQA-CONFLICT (1,840 episodes, *fully public, co-anchor*): derived from Wikidata entity revisions committed 2025-07-01 to 2025-12-15. Episodes pair pre-revision facts, post-revision corrections, and procedural templates. 31% of episodes contain direct contradictions; slice thresholds are $c \geq 0.74$ and $c \leq 0.19$ (n=460 each), with a $6{,}400$-decision training log.
- APIMIGRATE-PUBLIC (3,360 episodes, *fully public*): 96 natural changes across 18 API families, assembled from versioned documentation, release notes, schemas, and migration guides. Semantic documentation, procedural call recipes, and episodic executions form the three stores; no contradiction or poisoning is author-injected. API-disjoint train/validation/test splits use 10/3/5 families and 7,600 randomized training decisions. The test set has 560 conflict-heavy and 560 low-conflict episodes.

AGENTMEM-WILD (2,312 episodes) raises the store count to five, and TOOLTRACE-NAT (1,480 episodes) supplies naturally occurring migrations for the pre-specified placebo in Appendix C. Appendix B gives their trace windows, slice definitions, and logging budgets.

5.4 BASELINES

Retrieval controls. Unified Retriever applies revision-aware reranking to one pooled index; Mixture-of-Retrievers aggregates separate store retrievers; Heuristic Priority uses fixed ordering with staleness filtering; and Learned Router fits an MLP to observational outcomes.

Matched decision controls. EMSR (Exposure-Matched Supervised Router) shares TRUCE's randomized log, encoder, 3-layer 512-unit MLP capacity, optimizer, and latency mask, but minimizes observed-action binary cross-entropy and selects the feasible subset with highest predicted success, without AIPW correction, uncertainty gating, or abstention. LINUCB-CTX uses the identical feature pipeline and tunes α=0.6 on the same B=0.15 log. SUBSETPOLICY-DML, the direct-method learner, uses the shared MLP and features to minimize cross-fitted plug-in regret over all feasible subsets, without AIPW correction or abstention. SameLog-DR Policy instead fits eight action-value heads from action-level AIPW pseudo-outcomes and records an abstention before using Unified Retriever outside logged support. Both subset controls mask infeasible actions and tune only learning rate and weight decay on the validation split. Exact Score Search enumerates every feasible subset under TRUCE's fitted score without extra features, labels, or outcome queries.

Rule and upper-bound controls. ConflictRule selects the freshest non-contradicted store using the same metadata; Oracle selects stores from ground-truth labels. Together, EMSR isolates value beyond randomized logging, SUBSETPOLICY-DML controls the feature set, SameLog-DR Policy challenges complete subset decisions on the same interventional log, and Exact Score Search measures the greedy realization gap.

5.5 IMPLEMENTATION

Tool-agent executions use a Llama-3.1-70B backbone (Llama Team, AI @ Meta, 2024) served by vLLM 0.4.2 on 8×NVIDIA A100-80GB SXM4 GPUs (CUDA 12.4, driver 550.54.15, PyTorch 2.3.1); the router encoder is specified in Section 4.3. Unless noted, $B = 0.15$ and $\Lambda = 200$ ms. TOOLROUTING-BENCH uses five seeds; all public tables use an independent ten-seed block. Per-method entries are mean ± seed standard deviation. Inference for differences uses 10,000 paired

cluster-bootstrap resamples, grouping decisions by episode; APIMIGRATE-PUBLIC additionally resamples update events and preserves seed pairing. A prespecified family contains each public factual-revision, API-transfer, policy-layer, and component contrast once, with Benjamini–Hochberg FDR control at $q = 0.05$; attack analyses form a separate family.

6 RESULTS

We begin with the phenomenon rather than an aggregate leaderboard: *Q1* asks whether disagreement itself predicts routing value and survives falsification; *Q2* asks whether supported typed interactions turn that signal into successful calls across updates; *Q3* measures how much subset search and randomized exposure are needed to realize the fitted value.

6.1 Q1. DOES DISAGREEMENT IDENTIFY A ROUTING-VALUE REGIME?

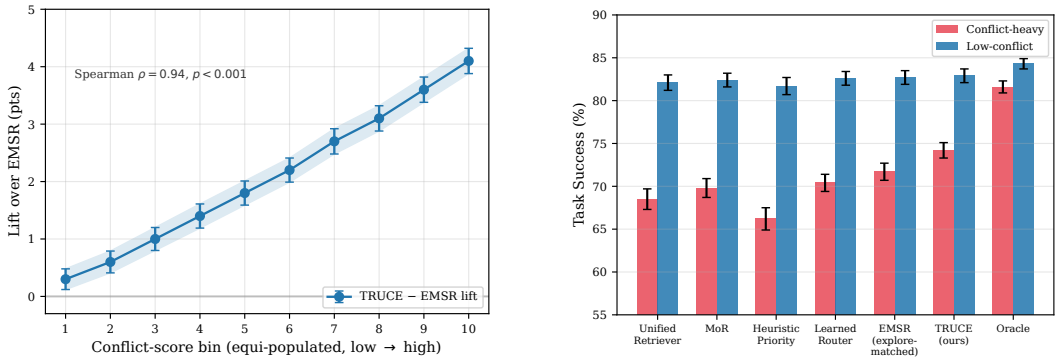

Figure 1: Conflict–value response. Left: pooled factual-revision and five-store evaluations rise from $+0.3$ to $+4.1$ TRUCE–EMSR points across ten equal-count conflict bins (ρ=0.94, p<0.001); error bars show 95% bootstrap intervals. Right: the controlled benchmark localizes the separation to conflict-heavy episodes. The predictor is measured before routing, and neither panel is defined by the selected action.

The ten-bin response separates conflict dependence from a uniform method advantage. The prespecified TOOLTRACE-NAT placebo then preserves retrieved content while permuting store identity: the TRUCE–Learned Router gap falls from $+3.9$ to $+0.12$ points (95% CI $[-0.41, +0.63]$, p=0.71), and outcome-invariant retrieval yields $+0.19$ points (p=0.48). The value signal therefore depends on typed, action-relevant disagreement rather than retrieval volume or a globally stronger predictor.

The same regime appears under two factual-revision constructions and natural API migrations:

Across 5,300 and 4,600 paired conflict-heavy episode-seed outcomes, TRUCE records 3493/5300 and 2985/4600 successes, compared with 3348/5300/2857/4600 for EMSR and 3352/5300/2880/4600 for SUBSETPOLICY-DML. Paired 95% CIs versus EMSR are [+1.6,+3.8]/[+1.5,+4.0]; versus SUBSETPOLICY-DML they are [+1.5,+3.9]/[+1.1,+3.5]. Low-conflict gaps are $+0.3/+0.4$ points, while stale failure falls by 2.8/2.7 and contradiction misuse by 2.3/2.3.

Table 1: Public factual-revision profile. Conflict-heavy success is mean $\pm$ seed SD over ten seeds; Δ rows are paired differences. Appendix A reports low-conflict utility, failure modes, and latency.

Method	WIKIREV-PUBLIC	EDITQA-CONFLICT
SubsetPolicy-DML	63.2 ± 1.0	62.6 ± 1.2
EMSR	63.2 ± 1.0	62.1 ± 1.2
TRUCE	65.9 ± 0.9	64.9 ± 1.1
Δ vs. EMSR	$+2.7$	$+2.8$
Δ vs. SubsetPolicy-DML	$+2.7$	$+2.3$

Table 2: Natural-API-update profile on API-family-disjoint test episodes. Success entries are mean ± seed SD over ten seeds; latency is p95 in ms.

Method	Conflict success	Low-conflict	p95
Unified Retriever	58.7 ± 1.4	77.5 ± 1.0	146
ConflictRule	60.8 ± 1.3	77.8 ± 0.9	138
LinUCB-Ctx	63.3 ± 1.2	78.0 ± 0.9	164
SubsetPolicy-DML	64.1 ± 1.1	78.1 ± 0.9	168
EMSR	63.7 ± 1.2	78.0 ± 0.9	166
SameLog-DR Policy	64.8 ± 1.1	78.1 ± 0.9	178
Exact Score Search	66.6 ± 1.0	78.2 ± 0.9	192
TRUCE	66.4 ± 1.0	78.2 ± 0.9	173

On APIMIGRATE-PUBLIC, TRUCE records 3718/5600 successes versus 3567/5600 for EMSR and 3590/5600 for SUBSETPOLICY-DML. Paired gains are +2.7 (95% CI [+1.5,+3.9], BH q=0.019) and +2.3 ([+1.1,+3.5], q=0.019). The same-log DR policy trails by 1.6 points ([+0.5,+2.7]). Stale-signature failure drops from 14.6 to 11.9, contradiction-induced misuse from 10.7 to 8.4, and low-conflict success changes by +0.2 points. Across all three public settings, task success and update-specific failures move together in the conflict-heavy regime.

6.2 Q2. WHICH SUPPORTED INTERACTIONS TURN CONFLICT INTO SUCCESSFUL CALLS?

Table 3: Controlled mechanism results on the held-out TOOLROUTING-BENCH test split. Entries are mean ± seed SD over five seeds; paired uncertainty applies to contrasts in text.

Method	Conflict success	Low-conflict	Stale fail	Misuse	p95 ms
Unified Retriever	68.5 ± 1.2	82.1 ± 0.9	14.3 ± 0.8	11.7 ± 0.7	142
Mixture-of-Retrievers	69.8 ± 1.1	82.4 ± 0.8	12.9 ± 0.9	10.4 ± 0.8	168
Heuristic Priority	66.2 ± 1.3	81.7 ± 1.0	11.8 ± 0.7	12.1 ± 0.9	131
Learned Router (no interv.)	70.4 ± 1.0	82.6 ± 0.8	11.2 ± 0.8	9.8 ± 0.7	155
EMSR (explore-matched, no DR)	71.7 ± 1.0	82.7 ± 0.8	10.3 ± 0.7	8.9 ± 0.6	159
TRUCE (ours)	74.2 ± 0.9	82.9 ± 0.8	8.9 ± 0.6	7.3 ± 0.5	163
Oracle	81.6 ± 0.7	84.3 ± 0.6	3.1 ± 0.4	2.8 ± 0.3	148

On conflict-heavy episodes, TRUCE gains $+2.5$ points over EMSR (p=0.006), reduces stale failure by 1.4 and misuse by 1.6, and remains at 163 ms p95. Across 790 paired episode-seed outcomes, successes are 586/790 versus 566/790; low-conflict successes are 659/795 versus 658/795.

Attribution and support. Replacing DR with the same-feature supervised head costs 2.9 points; removing randomized exposure costs 3.6, and removing pairwise scoring, confidence gating, or corroboration costs 0.8/1.1/1.8 on TOOLROUTING-BENCH. The public nested decomposition holds the feature pipeline, action-support mask, Unified fallback, and latency cap fixed. Before the veto, singleton-plus-pairwise DR reaches 64.8/63.8, $1.6/1.2$ points above the no-DR route; pair terms account for $0.7/0.6$ of that separation. Adding the common corroboration stage changes 14.6%/15.3% of conflict-heavy routes and adds $+1.1/+1.1$ points. The final-action support audit retains 93.6%/92.9% of public decisions, with empirical 95% pseudo-outcome prediction coverage of 92.7%/93.1%; abstention rises in the weakest-support conflict cells (Appendix H).

6.3 Q3. How much search and exploration realize the fitted value?

Table 4: Policy realization on public conflict-heavy episodes. Exact search enumerates every feasible subset under the same fitted score; agreement is with its selected action.

Method	WikiRev	EditQA	Agreement	p95 ms
SameLog-DR Policy	64.8	63.9	—	174/179
TRUCE	65.9	64.9	94.1/93.4%	167/171
Exact Score Search	66.1	65.1	100%	186/190

TRUCE exceeds the end-to-end same-log learner by +1.1/+1.0 points, with paired 95% CIs of [+0.2,+2.0]/[+0.1,+1.9]. Its 0.2/0.2-point gaps to exact fitted-score search have paired CIs of [-0.3,+0.7]/[-0.4,+0.8], both including zero. Greedy composition therefore captures nearly all value available to the fitted surrogate while saving 19 ms p95 on both benchmarks.

Cost saturation. The prespecified sufficiency rule—DR MAE at most 0.03, feasible-subset rank disagreement below 5%, and public lift within 0.3 points of the $B = 0.15$ reference—is first met at $B = 0.10$, where rank disagreement is 4.8%. At $B = 0.08$, the full stack attains $+2.34 \pm 0.34$ points, 85% of the $B = 0.15$ point-estimate lift, using 52% of its measured A100-GPU collection-and-training time.

7 Discussion

Conflict–value profiles separate two questions that aggregate tool-use accuracy conflates: whether memory disagreement predicts avoidable failures and whether a specialized route realizes enough conditional value to justify its statistical and systems cost. Benchmark-wise high/low contrasts establish the operating boundary; the graded response and type-identity intervention probe the predictor; matched-log, same-feature, and same-log controls locate value beyond extra data or metadata; and exact fitted-score search bounds the loss from greedy composition. The corresponding reductions in stale calls and contradiction misuse show that the success lift tracks the update failures that motivate the method.

The evidence is strongest for three typed stores under factual and API updates, with controlled scaling through $K = 8$ and explicit post-clipping support. Public decomposition assigns 0.6–0.7 points to pairwise scoring and 1.1 points to corroboration on the two factual tasks, so the reported end-to-end effect belongs to their composition rather than to the local estimands alone. The five held-out API families test family-disjoint migration behavior with update-event-clustered intervals; broader family-population inference and higher-order interactions beyond the observed $K = 9$ latency boundary are the natural next stress axes. Within the evaluated regime, support-aware fallback and the 200 ms cap keep the profile tied to a defined operating region.

8 Conclusion

Typed-memory conflict predicts the value of interventional routing after tool and knowledge updates. Conflict–value profiles expose that relation as conditional gain with explicit support, latency, and exploration cost, while TRUCE supplies a post-clipping identification-to-action implementation. Across public factual revisions and family-disjoint natural API migrations, the conflict-heavy advantage survives matched decision controls, low-conflict differences shrink to 0.2–0.4 points, and store-identity randomization removes the effect. Greedy routing then stays within 0.2 success points of exact fitted-score search on both public-core tasks. The resulting decision rule is concrete: reserve statistically and computationally intensive routing for update episodes where typed disagreement predicts material call-level reliability gains.

Reproducibility Statement

Sections 3–6 specify estimands, logging, models, splits, baselines, hardware, seeds, and uncertainty; the appendices provide proof details, per-benchmark results, overlap audits, ablations, robustness analyses, and operating-point sweeps.

AI Use Statement

Generative AI tools assisted manuscript preparation, including language and presentation editing, formatting, and bibliographic consistency checks. GPT-4o also served as the user simulator for the 1,200 custom episodes in the controlled ToolRouting-Bench component described in Appendix B; the two factual-revision benchmarks and APIMigrate-Public do not use that simulator. The authors take responsibility for the benchmark design, reported analyses, interpretations, and final text.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations*, 2024.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 2002.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning*, pp. 2206–2240, 2022.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

Nicola De Cao, Wilker Aziz, and Ivan Titov. Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6491–6506, 2021.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *International Conference on Machine Learning*, pp. 1097–1104, 2011.

Weiquan Huang, Zixuan Wang, Hehai Lin, Sudong Wang, Bo Xu, Qian Li, Beier Zhu, Linyi Yang, and Chengwei Qin. AMA: Adaptive memory via multi-agent collaboration. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 3099–3120, 2026. doi: 10.18653/v1/2026.findings-acl.152. URL https://aclanthology.org/2026.findings-acl.152/.

Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 874–880, 2021.

Dongming Jiang, Yi Li, Guanpeng Li, and Bingzhe Li. MAGMA: A multi-graph based agentic memory architecture for AI agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pp. 36848–36865, 2026. doi: 10.18653/v1/2026.acl-long.1709. URL https://aclanthology.org/2026.acl-long.1709/.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pp. 652–661, 2016.

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.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114 (13):3521–3526, 2017.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33: 9459–9474, 2020.

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, pp. 661–670, 2010.

Minghao Li, Yingxiu Zhao, Bowen Yu, Feifan Song, Hangyu Li, Haiyang Yu, Zhoujun Li, Fei Huang, and Yongbin Li. API-Bank: A comprehensive benchmark for tool-augmented LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 3102–3116, 2023.

Llama Team, AI @ Meta. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. doi: 10.48550/arXiv.2407.21783. URL https://arxiv.org/abs/2407.21783.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Kevin Meng, Arnab Sen Sharma, Alex J. Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *International Conference on Learning Representations*, 2023.

Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, et al. Augmented language models: A survey. *Transactions on Machine Learning Research*, 2023.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. Fast model editing at scale. In *International Conference on Learning Representations*, 2022.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

OpenAI. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

German I. Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019.

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

Daivik Patel and Shrenik Patel. ENGRAM: Effective, lightweight memory orchestration for conversational agents. *arXiv preprint arXiv:2511.12960*, 2025. doi: 10.48550/arXiv.2511.12960. URL https://arxiv.org/abs/2511.12960.

Shishir G Patil, Tianjun Zhang, Xin Wang, and Joseph E Gonzalez. Gorilla: Large language model connected with massive APIs. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/e4c61f578ff07830f5c37378dd3ecb0d-Abstract.html.

Hongjin Qian, Zheng Liu, Peitian Zhang, Kelong Mao, Defu Lian, Zhicheng Dou, and Tiejun Huang. MemoRAG: Boosting long context processing with global memory-enhanced retrieval augmentation. In *Proceedings of the ACM Web Conference 2025*, pp. 2366–2377, 2025. doi: 10.1145/3696410.3714805. URL https://doi.org/10.1145/3696410.3714805.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. ToolLLM: Facilitating large language models to master 16000+ real-world APIs. In *International Conference on Learning Representations*, 2024.

Peijun Qing, Xingjian Diao, Chiyu Ma, Saeed Hassanpour, and Soroush Vosoughi. Tailoring memory granularity for multi-hop reasoning over long contexts. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 3648–3666, 2026. doi: 10.18653/v1/2026.findings-eacl.189. URL https://aclanthology.org/2026.findings-eacl.189/.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/77c33e6a367922d003ff102ffb92b658-Abstract-Conference.html.

Yi Su, Maria Dimakopoulou, Akshay Krishnamurthy, and Miroslav Dudík. Doubly robust off-policy evaluation with shrinkage. In *International Conference on Machine Learning*, 2020.

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*, 2023.

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.

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 *International Conference on Learning Representations*, 2024.

A Detailed Public-Core Results

Tables 10 and 11 expand the main public-core summary with stale-failure, contradiction-misuse, low-conflict, and latency results. Both use disjoint evaluation episodes and the ten-seed protocol.

B Implementation Details

Pre-routing relation labels. The parser canonicalizes case, Unicode, numeric units, API aliases, and argument order before forming entity–property or API-family–endpoint–field keys. Equal-key items agree when normalized scalar payloads match or one schema is a backward-compatible refinement. They contradict when scalar values differ at the same revision slot, an argument is required in one record and removed in another, or types, enums, or ranges are disjoint; unresolved key matches and unequal keys are neutral. The newest task-relevant item is the lexicographic maximum of timestamp and revision identifier among retrieved records matching the request key. Missing timestamps contribute to staleness but not timestamp span; a remaining tie uses the fixed episodic–semantic–procedural order. Gold answers, execution outcomes, and post-route text are unavailable to this parser. The same labels enter D, pair features, and corroboration.

Benchmark protocol manifest. Table 5 places construction, grouping, training evidence, and held-out scoring in one ledger. Factual-revision groups are keyed jointly by entity, property, and revision event before either Wikidata-derived benchmark is partitioned; a shared group is forced to the same side of the train/evaluation boundary in both constructions. The two factual benchmarks are therefore treated as related task constructions, not as statistically independent samples. API migration events inherit their API-family group, so versions and endpoints from a held-out family cannot enter training.

Table 5: Benchmark protocol manifest. "High/low" gives the held-out conflict cohorts used in the primary profile.

Benchmark	Construction and target check	Grouping boundary	Training / held-out evaluation
TOOLROUTING-BENCH	API-Bank, ToolLLM, and custom multi-session tasks; success requires the executable subgoal assertion	Source episode and API identity	2,953/632/633 train/validation/test episodes; 12,000 logged train decisions; 158/159 high/low test
WIKIREV-PUBLIC	Wikidata revision chains; target is the normalized adjudicated post-revision payload	Entity–property–revision event, jointly with EditQA	2,120 episodes; 7,200 logged train decisions; 530/530 high/low held out
EDITQA-CONFLICT	Independently generated QA from Wikidata revision chains; target is checked against the committed revision payload	Entity–property–revision event, jointly with WikiRev	1,840 episodes; 6,400 logged train decisions; 460/460 high/low held out
APIMIGRATE-PUBLIC	Versioned documentation, release notes, schemas, and migration guides; calls must parse and satisfy post-migration assertions	API family; 10/3/5 families for train/validation/test	3,360 episodes and 96 events; 7,600 logged train decisions; 560/560 high/low test

Hardware and software. All experiments were conducted on a single node with 8×NVIDIA A100-80GB SXM4 GPUs (NVLink), AMD EPYC 7763 CPUs (128 cores, 2.45 GHz base), 1 TB DDR4 RAM, and 24 TB NVMe scratch. Software stack: CUDA 12.4 (driver 550.54.15), PyTorch 2.3.1 (+cu124), vLLM 0.4.2, Transformers 4.41.0, Flash-Attention 2.5.8, Python 3.11.7, NumPy 1.26.4, SciPy 1.13.1. Cross-fitting and paired-bootstrap computations use econml 0.15.0 and scikit-learn 1.4.2.

Total compute. The original five-benchmark study uses 948 A100-GPU-hours. APIMIGRATE-PUBLIC, SameLog-DR Policy, exact-search evaluation, public decomposition, and the adaptive holdout add 178 hours, for 1,126 A100-GPU-hours and 140.8 hours at eight-GPU data-parallel throughput. The six exploration logs contain 44,400 decisions. Pair scoring evaluates $\binom{3}{2} = 3$ terms on public tasks, 10 at $K = 5$, 21 at $K = 7$, and 28 at $K = 8$.

Benchmark construction. TOOLROUTING-BENCH filters API-Bank episodes to at least three calls, samples ToolLLM multi-tool tasks, and uses GPT-4o (OpenAI, 2024) as the user simulator for custom episodes. WIKIREV-PUBLIC and EDITQA-CONFLICT use 2025–2026 Wikidata revision diffs under CC-BY-SA-4.0. APIMIGRATE-PUBLIC identifies semantically versioned changes from public documentation, release notes, schemas, and migration guides, groups endpoints by API family before splitting, and converts each change into execution tasks whose three stores reflect only naturally published version differences. AGENTMEM-WILD is a five-store setting with 34% multi-store contradictions, $c \geq 0.76$ / $c \leq 0.21$ slices (n=578 each), and a $6{,}400$-decision log. TOOLTRACE-NAT spans a disjoint 45-day trace with 62 natural API migrations across 11 APIs, no injected contradictions, $c \geq 0.69$ / $c \leq 0.15$ slices (n=370 each), and a $4{,}800$-decision log. Both use anonymized aggregate commercial traces under operator data-use terms.

Leakage audit. We compare held-out episode text by exact 13-grams against the accessible API-Bank and ToolLLM training sets, a 2023-12-01 Wikidata dump, and public API-document snapshots dated through 2023-12-31. Match rates are 0.07%, 0.03%, 0.04%, and 0.09%; manual inspection attributes matches to short schema boilerplate and endpoint names. Every target revision or migration event occurs after 2025-07-01.

Table 6: Baseline fairness matrix. Neural methods share retrieval depth, reranker, latency cap, and context budget; matched decision controls also share randomized logs with TRUCE.

Method	Retrieval Depth	Reranker	Lat. Cap (ms)	Tok. Budget	Explore B	Random log
Unified Retriever	top-10	Cross-encoder	200	2048	—	no
Mixture-of-Retrievers	top-10 per store	Cross-encoder	200	2048	—	no
Heuristic Priority	top-10	None	200	2048	—	no
Learned Router (no interv.)	top-10	Cross-encoder	200	2048	—	no
EMSR (explore-matched)	top-10	Cross-encoder	200	2048	0.15	yes
SubsetPolicy-DML	top-10	Cross-encoder	200	2048	0.15	yes
SameLog-DR Policy	top-10	Cross-encoder	200	2048	0.15	yes
TRUCE (ours)	top-10	Cross-encoder	200	2048	0.15	yes

Corroboration heuristic. The deterministic step receives the selected stores, top retrieved items, timestamps, revision identifiers, confidence values, and pairwise entailment labels. It removes a store whose item is contradicted by a fresher selected item and lacks an agreeing partner. If none survive, it returns the store with the largest gated singleton score. It can delete stores or collapse to one store, never add a store.

Paired-bootstrap procedure. For a method pair (A, B) and metric m, we resample episodes with replacement $R = 10{,}000$ times at the episode level, preserving per-seed alignment: for each resample r we compute $\Delta_r = m(A)_r - m(B)_r$ averaged over seeds. The reported p-value is $\hat{p} = 2\min\{(1/R)\sum \mathbf{1}[\Delta_r \le 0], (1/R)\sum \mathbf{1}[\Delta_r \ge 0]\}$; confidence intervals use BCa correction with $z_0 = \Phi^{-1}((1/R)\sum \mathbf{1}[\Delta_r < \bar{\Delta}])$ and the standard jackknife acceleration.

Bounded-confounding sensitivity. For every supported background pair, define $e_{s|b}(x) = q(b \cup \{s\} \mid x, E{=}1)/\{q(b \cup \{s\} \mid x, E{=}1) + q(b \mid x, E{=}1)\}$. We bound hidden within-pair assignment odds by $e^{\star}_{s|b}(x,u)/(1-e^{\star}_{s|b}(x,u)) \le \Gamma\, e_{s|b}(x)/(1-e_{s|b}(x))$ and recompute worst-case background-averaged AIPW effects across the Γ sweep.

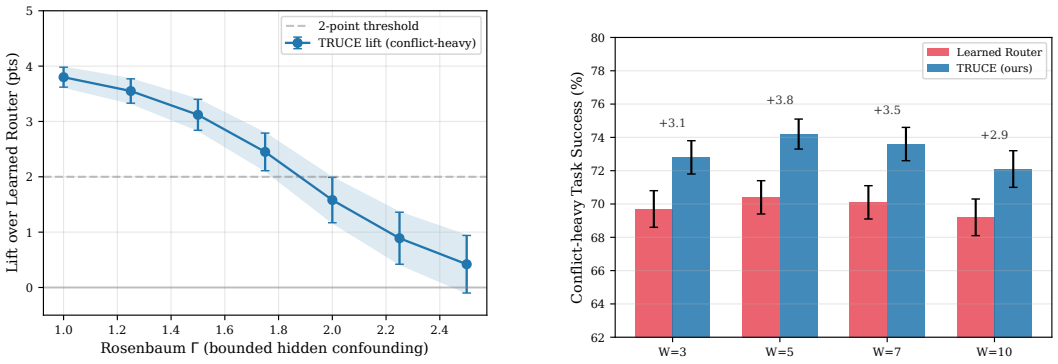

Figure 2: Left: bounded hidden-confounding sensitivity. The plotted curve is the conflict-heavy task-success lift of TRUCE over LEARNED ROUTER as Rosenbaum Γ grows. The lift remains positive through the full sweep and stays above 2.4 points until roughly $\Gamma \approx 1.8$. Right: decision-window sweep. The paired lift is positive for every $W \in \{3, 5, 7, 10\}$ and peaks at $W{=}5$. The two panels therefore stress-test the two assumptions that matter most for the paper's local causal reading.

Table 7: Component ablation on TOOLROUTING-BENCH conflict-heavy. Entries are mean ± seed SD over five seeds. Every row holds the log, encoder, backbone, features, and latency cap fixed.

Variant	Task Success ↑	Stale Fail ↓	Calibration ECE ↓
Full TRUCE (additive+pairwise)	74.2 ± 0.9	8.9 ± 0.6	0.041
− Pairwise interaction term $\hat{\phi}$	73.4 ± 0.9	9.4 ± 0.6	0.046
− Cross-fitting	72.8 ± 1.0	9.7 ± 0.7	0.068
− Randomized exposure	70.6 ± 1.1	11.0 ± 0.8	0.092
− Uncertainty gating	73.1 ± 0.9	9.2 ± 0.7	0.057
− Corroboration step	72.4 ± 1.0	10.1 ± 0.7	0.044
IPS-only (no outcome model)	71.3 ± 1.2	10.8 ± 0.9	0.103

Table 8: Robustness at 10% injection across four attack families. Entries are mean ± seed SD over five seeds. Local retain measures neighboring non-target facts; the final column reports relative TRUCE reduction versus Learned Router/EMSR.

Attack Family	Attack Success ↓				Local Retain ↑		Rel. Δ
	Unified	Learned Rtr.	EMSR	TRUCE	Learned Rtr.	TRUCE	(vs L/E, %)
Direct Alias	24.3 ± 1.1	20.1 ± 1.0	18.5 ± 0.9	14.8 ± 0.8	77.9 ± 0.7	77.6 ± 0.7	-26.4 -20.0
Paraphrase	21.7 ± 1.0	18.4 ± 0.9	16.8 ± 0.9	14.3 ± 0.8	78.8 ± 0.7	78.4 ± 0.7	-22.3 -14.9
Delayed Revision	29.6 ± 1.3	25.8 ± 1.2	23.6 ± 1.1	16.1 ± 0.9	76.6 ± 0.8	76.2 ± 0.8	-37.6 -31.8
Corpus Churn	26.1 ± 1.1	22.3 ± 1.0	20.6 ± 1.0	15.4 ± 0.8	77.1 ± 0.8	76.8 ± 0.7	-30.9 -25.2

Fairness controls. Shared retrieval and inference resources rule out depth, reranking, context budget, and latency as explanations for the gain. EMSR, SubsetPolicy-DML, and SameLog-DR additionally share randomized exposure; Section 5 gives their objectives, action masks, tuning rule, and fallback.

Ablation and robustness. Randomized exposure produces the largest component drop, while pairwise scoring, cross-fitting, gating, and corroboration each add success or calibration value. Across attacks, TRUCE lowers attack success by 14.9–31.8% relative to matched-log EMSR; local and global retain move by at most 0.5 points.

Support-aware behavior. Abstention rises as overlap degrades, aligning routed coverage with the support audit. The attack panel shows that robustness remains after randomized logging is held fixed through EMSR.

C PRE-SPECIFIED PLACEBO ON TOOLTRACE-NAT

Identity-intervention protocol. The randomization unit is the episode–seed pair. One of the six bijections of the three store identities is drawn uniformly; identity draws are retained. The bijection is applied before feature extraction and model fitting to the store-name channels only, while item content, rank, timestamps, revision identifiers, actions, outcomes, split membership, and the original pre-routing conflict slices remain fixed. TRUCE, EMSR, and Learned Router are each refit on their correspondingly relabeled training log with the original architecture, tuning grid, and seed, then evaluated on paired episodes under the same bijection. A second intervention changes only the target to paraphrase identification, whose outcome is invariant to store identity. This protocol distinguishes reliance on typed source identity from gains due to content volume or randomized logging.

Table 9: Global retain quality under the four attacks. Entries are mean ± seed SD over five seeds. Retain remains nearly constant while TRUCE reduces attack success in Table 8.

Attack Family	Unified Ret.	Learned Router	EMSR	TRUCE
Direct Alias	80.4 ± 0.6	80.1 ± 0.6	80.0 ± 0.6	79.8 ± 0.6
Paraphrase	81.1 ± 0.6	80.9 ± 0.6	80.7 ± 0.6	80.6 ± 0.6
Delayed Revision	79.7 ± 0.7	79.4 ± 0.6	79.3 ± 0.6	79.1 ± 0.6
Corpus Churn	80.0 ± 0.6	79.6 ± 0.6	79.5 ± 0.6	79.3 ± 0.6

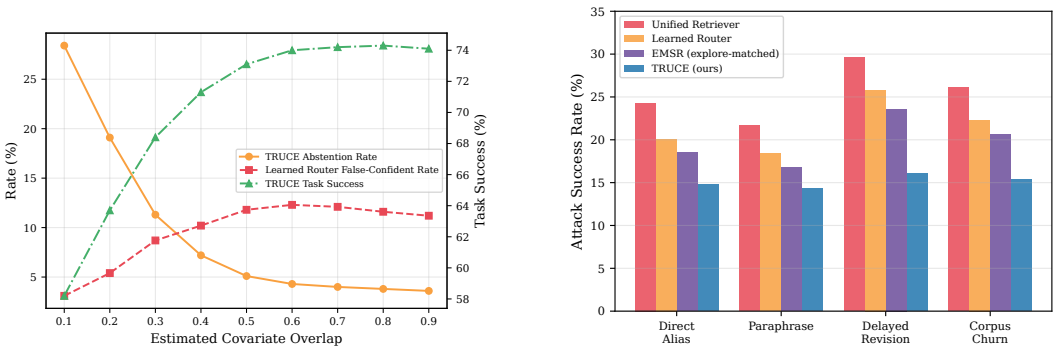

Figure 3: Left: abstention as a function of estimated covariate overlap. TRUCE abstains increasingly as overlap degrades, which is exactly the regime where causal estimates become least reliable. LEARNED ROUTER remains confident in the same region and pays for it in stale failure. Right: attack success under the four 10% injection attacks. The EMSR bar separates the benefit of randomized logging from the benefit of DR estimation. Together, the panels tie caution under weak identification to measurable robustness under adversarial stress.

Results. Under identity randomization, TRUCE's conflict-heavy advantage over Learned Router changes from $+3.9$ to $+0.12$ points, a paired shrinkage of 3.78 points; the post-intervention gap has 95% CI $[-0.41, +0.63]$ (p=0.71, 10,000 episode-paired resamples). Its advantage over EMSR changes from $+2.4$ to $+0.07$ points (p=0.83). Under the outcome-invariant target, the TRUCE–Learned Router gap is $+0.19$ points ($[-0.31, +0.69]$, p=0.48). Typed identity and action relevance are therefore both necessary for the measured conflict advantage.

D PUBLIC-DATA BENCHMARK: WIKIREV-PUBLIC

WIKIREV-PUBLIC tests factual updates through revision-resolution QA built from Wikidata entity diffs under CC-BY-SA-4.0. EDITQA-CONFLICT provides a second factual-revision task construction, while APIMIGRATE-PUBLIC supplies the complementary natural API-update setting.

TRUCE gains $+2.7$ over both EMSR (p=0.007, BH q=0.019) and SUBSETPOLICY-DML. Success, stale failure, and misuse improve together, while learned methods remain within 0.5 points in low conflict.

E PUBLIC-DATA CO-ANCHOR: EDITQA-CONFLICT

EDITQA-CONFLICT is built end-to-end from Wikidata entity-revision diffs under CC-BY-SA-4.0 and tests a second task construction within factual revisions.

Reading Table 11. On EDITQA-CONFLICT, TRUCE gains $+2.8$ over EMSR (p=0.009, BH q=0.019) and $+2.3$ over SUBSETPOLICY-DML. Stale failure and contradiction misuse fall by 2.7 and 2.3 points relative to EMSR, while the low-conflict gap is $+0.4$.

Table 10: WIKIREV-PUBLIC results. Entries are mean ± seed SD over ten seeds; paired intervals apply to method differences reported in Section 6.1.

Method	Conflict-heavy ↑	Low-conflict ↑	Stale Fail ↓	Misuse ↓	Lat. p95 (ms)
Unified Retriever	58.1 ± 1.2	73.2 ± 0.9	21.3 ± 1.0	14.2 ± 0.8	139
Mixture-of-Retrievers	59.4 ± 1.1	73.6 ± 0.9	19.6 ± 0.9	13.1 ± 0.8	166
Heuristic Priority	57.3 ± 1.2	72.8 ± 1.0	18.2 ± 0.9	14.8 ± 0.9	128
ConflictRule (deterministic)	60.9 ± 1.1	73.4 ± 0.9	15.1 ± 0.8	12.0 ± 0.7	134
Learned Router (no interv.)	61.8 ± 1.0	74.1 ± 0.8	14.6 ± 0.8	11.3 ± 0.7	153
LinUCB-Ctx (bandit)	62.9 ± 1.0	74.2 ± 0.8	13.7 ± 0.7	10.6 ± 0.7	158
SubsetPolicy-DML	63.2 ± 1.0	74.3 ± 0.8	13.1 ± 0.7	10.1 ± 0.7	161
EMSR (explore-matched)	63.2 ± 1.0	74.3 ± 0.8	13.2 ± 0.7	10.1 ± 0.7	160
TRUCE (ours)	65.9 ± 0.9	74.6 ± 0.8	10.4 ± 0.6	7.8 ± 0.5	167
Oracle	72.3 ± 0.7	76.4 ± 0.6	4.2 ± 0.3	3.1 ± 0.3	150

Table 11: EDITQA-CONFLICT results under the same protocol as Table 10. Entries are mean ± seed SD over ten seeds.

Method	Conflict-heavy ↑	Low-conflict ↑	Stale Fail ↓	Misuse ↓	Lat. p95 (ms)
Unified Retriever	57.2 ± 1.5	73.8 ± 1.2	22.1 ± 1.1	13.8 ± 0.9	143
Mixture-of-Retrievers	59.1 ± 1.4	74.2 ± 1.1	20.4 ± 1.0	12.4 ± 0.8	170
Heuristic Priority	56.4 ± 1.4	73.1 ± 1.1	18.9 ± 1.0	14.1 ± 0.9	132
ConflictRule (deterministic)	60.2 ± 1.3	74.0 ± 1.0	15.7 ± 0.9	11.7 ± 0.8	138
Learned Router (no interv.)	60.6 ± 1.3	74.9 ± 1.0	15.1 ± 0.9	11.0 ± 0.7	156
LinUCB-Ctx (bandit)	62.0 ± 1.2	75.0 ± 1.0	14.2 ± 0.8	10.3 ± 0.7	162
SubsetPolicy-DML	62.6 ± 1.2	75.1 ± 1.0	13.6 ± 0.8	9.7 ± 0.7	165
EMSR (explore-matched)	62.1 ± 1.2	75.1 ± 1.0	13.9 ± 0.8	9.9 ± 0.7	163
TRUCE (ours)	64.9 ± 1.1	75.5 ± 1.0	11.2 ± 0.7	7.6 ± 0.6	171
Oracle	70.5 ± 0.9	77.1 ± 0.8	5.3 ± 0.4	3.4 ± 0.3	158

F APIMIGRATE-PUBLIC PROTOCOL AND FULL METRICS

The benchmark covers 18 public API families and 96 release events. Families, rather than endpoints, are assigned 10/3/5 to train/validation/test, preventing related versions from crossing splits. Each task exposes versioned documentation, a procedural call template, and timestamped execution traces. The primary metric is end-to-end task success; secondary metrics are stale-signature failure, contradiction-induced misuse, routing abstention, expected calibration error (ECE), latency, and exploration-plus-training cost. Ten seeds share the same test episodes. Confidence intervals cluster by update event and preserve method/seed pairing.

Route abstention is concentrated in the highest-disagreement, weakest-overlap cells. Exact search improves success by 0.2 points while adding 19 ms at p95. Relative exploration-plus-training costs for EMSR, SubsetPolicy-DML, SameLog-DR Policy, Exact Score Search, and TRUCE are 0.72/0.81/0.91/1.12/1.00.

Table 12: Full APIMIGRATE-PUBLIC conflict-heavy results. Success, failure, misuse, and abstention are percentages; entries before latency are mean ± seed SD where shown. Cost is relative to full TRUCE.

Method	Success	Stale	Misuse	Abstain	ECE	p50/p95 ms
EMSR	63.7 ± 1.2	14.6	10.7	0.0	0.071	92/166
SubsetPolicy-DML	64.1 ± 1.1	14.1	10.2	0.0	0.066	96/168
SameLog-DR Policy	64.8 ± 1.1	13.2	9.6	3.8	0.058	103/178
Exact Score Search	66.6 ± 1.0	11.7	8.2	6.7	0.044	113/192
TRUCE	66.4 ± 1.0	11.9	8.4	6.5	0.046	101/173

Table 13: Public-core decomposition on conflict-heavy episodes. EMSR is the native-policy reference; the rows below it are nested variants with a common feature pipeline, final-action support mask, Unified fallback, and 200 ms cap. The first three nested variants omit corroboration, which the full row adds. Each cell reports WikiRev/EditQA over ten seeds.

Variant	Success	Stale fail	Misuse	ECE	p95 ms
EMSR	63.2/62.1	13.2/13.9	10.1/9.9	.068/.071	160/163
Same features, no DR	63.2/62.6	13.1/13.6	10.1/9.7	.064/.067	161/165
DR singleton + greedy	64.1/63.2	12.1/12.8	9.1/8.9	.054/.058	158/162
DR singleton+pairwise, no corroboration	64.8/63.8	11.3/12.1	8.5/8.3	.048/.051	164/168
Full TRUCE	65.9/64.9	10.4/11.2	7.8/7.6	.043/.046	167/171

G PUBLIC-CORE DEPLOYED-STACK DECOMPOSITION

Before corroboration, replacing the no-DR route by singleton DR adds $+0.9/+0.6$ points and pairwise scoring adds another +0.7/+0.6 points (paired 95% CI [+0.1,+1.3]/[0.0,+1.2], BH q=$0.020/0.020$). Corroboration then changes 14.6%/15.3% of selected conflict-heavy subsets and adds +1.1/+1.1 points ([+0.4,+1.8]/[+0.3,+1.9], q=$0.019/0.020$). Because support and fallback are fixed across the nested rows, the table separates score estimation, pair composition, and the deterministic veto.

H CONDITIONAL OVERLAP AND SCORE CALIBRATION

The audit evaluates the final action after latency clipping: the required-action column is computed from the enumerated subset propensity, not a product of marginal floors. TRUCE routes only retained cells and falls back to Unified Retriever elsewhere. Coverage evaluates the routing intervals against held-out pseudo-outcomes; it does not relabel them as pointwise-effect confidence intervals.

Table 14: Final-action support by disagreement severity. Marginal p05 is the fifth-percentile store exposure; required-action p05 is the fifth percentile of the smallest logged $q(a \mid x, E{=}1)$ among the two or four final actions entering a retained contrast. Support is routed coverage; the last column is empirical 95% pseudo-outcome prediction coverage.

Benchmark	Severity	Marginal p05	Required-action p05	Support	Pseudo-outcome cov.
WikiRev	Low	.132	.021	96.8%	94.3%
WikiRev	All	0.116	0.016	93.6%	92.7%
WikiRev	Conflict-heavy	.104	.013	90.4%	91.1%
EditQA	Low	.129	.020	96.2%	94.0%
EditQA	All	0.113	0.015	92.9%	93.1%
EditQA	Conflict-heavy	.102	.012	89.9%	91.4%

I Exploration Budget as a Decision Threshold

Table 15: Validation calibration of exploration budget. Rank disagreement is the fraction of contexts whose top latency-feasible subset differs from the $B = 0.15$ reference; retained lift is the pooled public-core gain over matched-B EMSR.

B	DR MAE	Rank disagreement	Public lift
.02	.081	18.4%	+0.9
.04	.061	13.2%	+1.63
.08	.036	7.1%	+2.34
.10	.029	4.8%	+2.48
.12	.026	4.2%	+2.58
.15	.024	3.7%	+2.75
.20	.023	3.5%	+2.81
.25	.022	3.4%	+2.79

The prespecified sufficiency rule requires DR MAE at most 0.03, rank disagreement below 5%, and public lift within 0.3 points of the $B = 0.15$ reference. The first budget satisfying all three is $B = 0.10$; $B = 0.08$ remains the cost-efficient deployment point when a small lift reduction is acceptable.

J Threat, Utility, and Evidence Matrix

TRUCE reduces adaptive ASR by 5.9 points versus EMSR and 5.0 versus SameLog-DR Policy. Benign success changes by -0.4 points (95% CI [-0.9,+0.1]), preserving ordinary utility while improving held-out attack resistance.

K Public-Data Operating-Point Pareto

We retrain every configuration in $\{B \in \{0.04, 0.08, 0.12, 0.15, 0.20\}$, pairwise $\in \{$off,on$\}$, gating $\in \{$off,on$\}\}$ and evaluate pooled public-core conflict-heavy success. Cost includes exploration and nuisance training; 100% denotes the $B = 0.15$ full-stack reference.

Cost tradeoff. Pairwise scoring and uncertainty gating account for 29% and 12% of public lift, respectively; removing both loses 45%. The $B = 0.08$ full stack reaches $+2.34 \pm 0.34$ points (85% of the reference point estimate) at 52% normalized A100-GPU cost, while $B = 0.20$ adds only 0.06 points over the reference at 38% greater cost.

Table 16: Systems/security evidence matrix. Each row links the protected routing object to adversary capability, enforcement boundary, strongest comparator, utility cost, observed stress boundary, and manuscript evidence.

Protected object	Adversary and trust assumptions	Property and comparator	Held-out setting, utility, failure, evidence
Selected memory subset	Single-store content write; logs and revision IDs trusted	Reject stale or contradicted evidence; EMSR	Direct Alias/Paraphrase; retain within 0.3 points; coordinated writes exceed boundary; Tables 8–9
Post-update tool call	Delayed update or corpus churn; uncompromised stores trusted	Reduce update-induced misuse; SubsetPolicy-DML	Delayed Revision/Corpus Churn; p95 below 200 ms; compromised revision metadata defeats recency; Table 8
Adaptive route	Wording/timing adapts after observing public behavior; randomized logs immutable	Limit single-store attack success; SameLog-DR Policy	Attack and API families held out; -0.4-point benign change; multi-store corruption remains; Table 17

Table 17: Held-out adaptive attack over 840 episodes and ten seeds. Attack success rate (ASR) is lower-is-better; raw counts aggregate paired episode-seed outcomes.

Method	ASR	Raw ASR	p95 ms
EMSR	23.8%	1999/8400	169
SubsetPolicy-DML	22.5%	1890/8400	171
SameLog-DR Policy	22.9%	1924/8400	181
TRUCE	17.9%	1504/8400	176

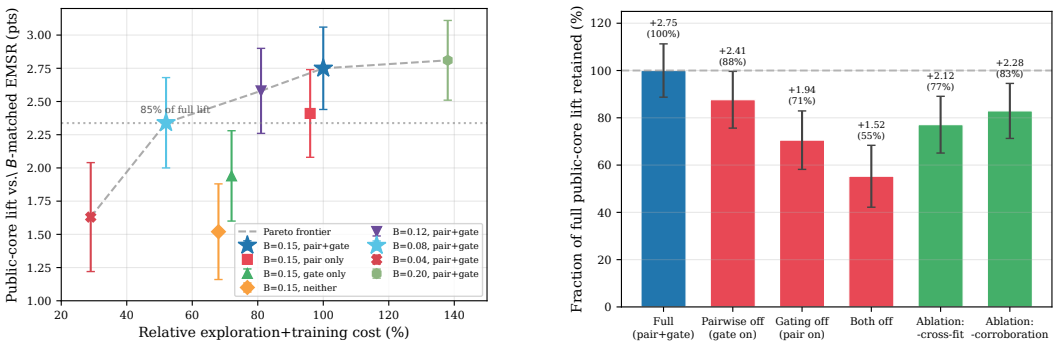

Figure 4: Left: public-core Pareto frontier over $\{B, \text{pairwise}, \text{uncertainty gating}\}$. The B=0.08 configuration with both components lies near the knee of the cost–lift curve. Right: public-core component attribution: pairwise (-29% of lift when off), uncertainty gating (-12%), and both jointly (-45%). Both panels use the pooled WIKIREV-PUBLIC∪EDITQA-CONFLICT evaluation set and connect routing value to its statistical and computational cost.

Table 18: Public-core operating-point Pareto on 990 conflict-heavy episodes and ten seeds. Public lift compares TRUCE with an EMSR router retrained at the same B. The $B = 0.08$ full-stack setting retains 85% of reference lift at 52% cost.

B	Pairwise	Uncert. gating	Public lift vs. B-matched EMSR ↑	Fraction of full lift	Rel. cost ↓
0.15	on	on	$+2.75 \pm 0.31$	100%	100%
0.15	on	off	$+2.41 \pm 0.33$	88%	96%
0.15	off	on	$+1.94 \pm 0.34$	71%	72%
0.15	off	off	$+1.52 \pm 0.36$	55%	68%
0.12	on	on	$+2.58 \pm 0.32$	94%	81%
0.08	on	on	$+2.34 \pm 0.34$	85%	52%
0.04	on	on	$+1.63 \pm 0.41$	59%	29%
0.20	on	on	$+2.81 \pm 0.30$	102%	138%

Table 19: Conflict-heavy transfer against same-feature baselines. WRP, TRB, EQ, AMW, and TTN denote WIKIREV-PUBLIC, TOOLROUTING-BENCH, EDITQA-CONFLICT, AGENTMEM-WILD, and TOOLTRACE-NAT; TOOLROUTING-BENCH uses five seeds and the other columns use ten. Δ compares TRUCE with the strongest alternative policy, SUBSETPOLICY-DML, for each benchmark.

Method	WRP	TRB	EQ	AMW	TTN
ConflictRule (deterministic)	60.9 ± 1.1	69.8 ± 1.2	60.2 ± 1.3	57.6 ± 1.4	59.1 ± 1.3
LinUCB-Ctx (bandit)	62.9 ± 1.0	71.9 ± 1.0	62.0 ± 1.2	59.1 ± 1.2	61.3 ± 1.2
SubsetPolicy-DML	63.2 ± 1.0	72.4 ± 1.0	62.6 ± 1.2	59.6 ± 1.2	61.8 ± 1.2
EMSR (explore-matched)	63.2 ± 1.0	71.7 ± 1.0	62.1 ± 1.2	58.9 ± 1.2	61.7 ± 1.2
TRUCE (ours)	65.9 ± 0.9	74.2 ± 0.9	64.9 ± 1.1	62.0 ± 1.1	64.1 ± 1.1
Δ vs. SubsetPolicy-DML	+2.7	+1.8	+2.3	+2.4	+2.3
paired p/q	0.007 / 0.013	0.014 / 0.014	0.009 / 0.013	0.008 / 0.013	0.010 / 0.013

L STRONG-BASELINE TRANSFER TABLE

Transfer. The same comparator rule selects SUBSETPOLICY-DML on all five benchmarks, and TRUCE remains ahead by 1.8–2.7 points. EMSR separately controls randomized data collection.

M FEATURE-FAMILY ABLATION

Feature attribution. Individual metadata removals cost 0.4–1.2 points. The supervised same-feature head loses 2.9 points and IPS loses 2.3, showing that neither one metadata family nor weighting alone reproduces the full result.

N AGENT-SPECIFIC IDENTIFICATION: PROOF SKETCH AND EMPIRICAL CHECKS

Exact clipping probability. Store s receives validation-measured planning cost ℓ_s, and a subset is feasible when $\sum_{s \in a} \ell_s \leq \Lambda$. Let $\mathcal{R}(p, a)$ contain removal paths $r_0 = p, \ldots, r_m = a$ that remove one uniformly sampled store at each step and whose intermediate sets are infeasible. With $h(p \mid p) = 1$

Table 20: Feature-family ablation on TOOLROUTING-BENCH conflict-heavy. Entries are mean ± seed SD over five seeds. Rows preserve DR, cross-fitting, gating, and corroboration unless the estimator is named explicitly.

Feature configuration	Task Success ↑	Δ vs. full	Stale Fail ↓
Full TRUCE features + DR	74.2 ± 0.9	—	8.9 ± 0.6
− contradiction flags	73.0 ± 1.0	-1.2	9.8 ± 0.7
− revision-recency features	73.3 ± 1.0	-0.9	9.6 ± 0.7
− entailment-derived pair signals	73.5 ± 0.9	-0.7	9.4 ± 0.7
− staleness indicators	73.6 ± 1.0	-0.6	9.3 ± 0.7
− task-phase proxies	73.8 ± 0.9	-0.4	9.1 ± 0.7
Same features, no DR (supervised MLP on $\hat{\mu}$)	71.3 ± 1.0	-2.9	10.7 ± 0.8
Same features, IPS-only	71.9 ± 1.1	-2.3	10.4 ± 0.8

for feasible proposals,

$$h(a \mid p) = \sum_{r \in \mathcal{R}(p,a)} \prod_{j=0}^{m-1} |r_j|^{-1}, \qquad q(a \mid x, E{=}1) = 2^{-|\mathcal{S}(x)|} \sum_{p \sqsupseteq a} h(a \mid p).$$

The logger enumerates this finite sum and records it with the final action. End-to-end measured latency, rather than the additive planning surrogate, is reported in results.

Proof sketch of Proposition 1. For a supported final action a, define

$$\varphi_a(O; x) = \mu_a(x) + \frac{\mathbf{1}[A = a]}{q(a \mid x, E{=}1)}\{Y - \mu_a(x)\}.$$

Let $\delta_{s,b}(x) = m_{b\cup\{s\}}(x) - m_b(x)$. Centering both the conditional and marginal targets gives the singleton influence function

$$\mathrm{IF}_s(O) = \tau_s(X) - \tau_s + \sum_b g_s(b \mid X)\{\varphi_{b\cup\{s\}}(O; X) - \varphi_b(O; X) - \delta_{s,b}(X)\}.$$

The pair influence function replaces the two-action bracket by the corresponding centered four-action signed sum and g_s by $g_{ss'}$. In particular, the conditional residual variance for a fixed singleton background contains the action-specific terms

$$\frac{\mathrm{Var}(Y \mid X, A = b \cup \{s\}, E = 1)}{q(b \cup \{s\} \mid X, E = 1)} + \frac{\mathrm{Var}(Y \mid X, A = b, E = 1)}{q(b \mid X, E = 1)},$$

rather than a pooled $e(X)\{1 - e(X)\}$ denominator. Cross-fitting groups episodes, and empirical influence contributions are summed within episode before variance estimation. With independent episodes, bounded inverse probabilities on retained support, and the nuisance-rate condition of Chernozhukov et al. (2018, Thm. 3.1), a cluster central limit theorem gives asymptotic normality. Context-level scores use the separately calibrated routing intervals in Table 14; the averaged-parameter limit is not transferred to pointwise effects.

Windowed assignment. The continuation ρ is the frozen history-dependent routing law used after a focal exposure; it maps each post-exposure history to the next route, including its fixed randomization distribution. Thus $Y_{et}(a; \rho)$ integrates later routing generated under the same law for every a, and any effect mediated through those routes belongs to the focal total effect. Overlapping decisions may separately serve as focal observations, but they are not treated as independent: folds and variance calculations group the complete episode. On TOOLTRACE-NAT, 96.4% of routing-to-outcome arcs close within five steps and 98.7% within seven; Figure 5 reports the locality distribution.

Dose response. The left panel is a descriptive graded-response diagnostic over WIKIREV-PUBLIC, EDITQA-CONFLICT, and AGENTMEM-WILD; benchmark-wise high/low contrasts, including APIMIGRATE-PUBLIC, carry the cross-setting claim.

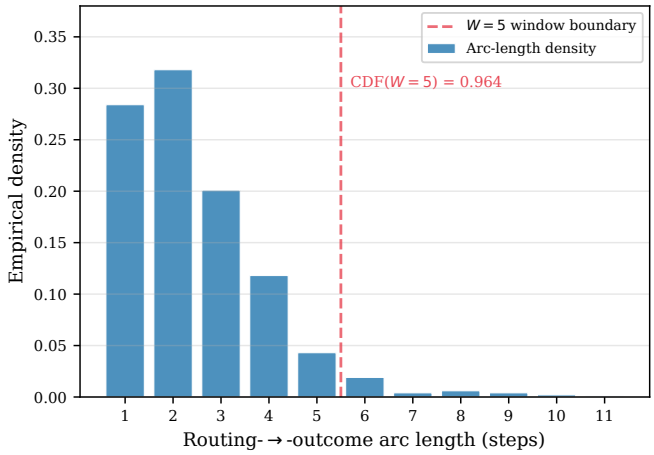

Figure 5: Routing-to-outcome locality on TOOLTRACE-NAT. Arcs record the number of tool steps between a routed memory exposure and its attributed outcome; 96.4% close by $W = 5$ and 98.7% by $W = 7$. The concentration supports the five-step assignment window used by the local exposure analysis.

O CONTINUOUS-CONFLICT AND MULTIPLICITY CONTROL

Continuous-conflict binning. Ten equi-populated bins over 6,272 pooled evaluation episodes contain 627–628 episodes each. From low to high conflict, TRUCE–EMSR lifts (bootstrap SE) are $0.3(.18)$, $0.6(.19)$, $1.0(.20)$, $1.4(.21)$, $1.8(.21)$, $2.2(.21)$, $2.7(.22)$, $3.1(.22)$, $3.6(.22)$, and $4.1(.22)$ points. Spearman $\rho = 0.94$ is computed over these ten bin means, and the endpoint difference has $p < 0.001$. Because the pool excludes APIMIGRATE-PUBLIC, Figure 1 is used as a dose-response diagnostic rather than the evidence for API-family transfer.

Multiplicity control. The central confirmatory family contains 13 prespecified contrasts: four public factual-revision comparisons, three API-transfer comparisons, two policy-layer comparisons, and four public component comparisons. Table 21 lists every member once; standard Benjamini–Hochberg adjustment gives a largest $q = 0.020$. Attack comparisons form a separate 11-contrast family (four attacks against two controls plus the adaptive attack against three controls), whose largest adjusted value is 0.024. Remaining sensitivity analyses are descriptive.

Table 21: Complete 13-contrast central confirmatory family with raw p and Benjamini–Hochberg q. All effects retain significance at FDR 0.05; the largest q is 0.020.

Contrast	raw p	BH q
TRUCE vs. EMSR, WikiRev	0.007	0.019
TRUCE vs. SubsetPolicy-DML, WikiRev	0.007	0.019
TRUCE vs. EMSR, EditQA	0.009	0.019
TRUCE vs. SubsetPolicy-DML, EditQA	0.009	0.019
TRUCE vs. EMSR, APIMigrate	0.006	0.019
TRUCE vs. SubsetPolicy-DML, APIMigrate	0.010	0.019
TRUCE vs. SameLog-DR Policy, APIMigrate	0.017	0.020
TRUCE vs. SameLog-DR Policy, WikiRev	0.015	0.020
TRUCE vs. SameLog-DR Policy, EditQA	0.019	0.020
Pairwise vs. singleton, WikiRev	0.014	0.020
Pairwise vs. singleton, EditQA	0.020	0.020
Full vs. no corroboration, WikiRev	0.009	0.019
Full vs. no corroboration, EditQA	0.012	0.020

Table 22: Same-construction scaling on AGENTMEM-WILD$^+$. Entries are mean $\pm$ seed SD over ten seeds. Latency remains below 200 ms through $K = 8$ as pair scores grow quadratically.

K	TRUCE ↑	EMSR	Δ vs. EMSR	Lat. p95 (ms)	#Pair scores
5 (AGENTMEM-WILD)	62.0 ± 1.1	58.9 ± 1.2	$+3.1$	174	10
7 (AGENTMEM-WILD$^+$)	60.3 ± 1.2	58.2 ± 1.3	$+2.1$	178	21
8 (AGENTMEM-WILD$^+$)	58.4 ± 1.3	56.7 ± 1.4	$+1.7$	186	28
9 (stress test, $> \Lambda$)	55.7 ± 1.5	54.9 ± 1.6	$+0.8$	214	36

Confirmatory family. The family tests transfer, comparator strength, policy approximation, and component attribution without reusing a contrast across categories. All adjusted values lie between 0.019 and 0.020 after rounding to three decimals.

P SCALING TO $K>5$: AGENTMEM-WILD$^+$

AGENTMEM-WILD$^+$ extends the same benchmark construction with long-horizon reflection and preference stores at $K = 7$, then an API-migration ledger at $K = 8$. Every other construction parameter and split is preserved.

Scaling. Lift remains 1.7 points at $K = 8$ with 186 ms p95 latency. At $K = 9$, 36 pair scores push p95 to 214 ms and reduce the gain to 0.8 points, locating the practical boundary of exhaustive pair scoring.

Q QUALITATIVE CASE STUDIES

Three cases illustrate pairwise prevention, a metadata-rule error, and a residual higher-order conflict.

Case 1 (pairwise prevention, AGENTMEM-WILD). Episode `AMW-1137` asks an agent for a project status update. The semantic store contains a high-confidence but stale 2025-09 project-milestone entry; the procedural store contains a fresh 2025-10 skill template that *requires* the stale milestone as a precondition. Singleton $\hat{\tau}$ scores both stores positive ($+0.18$, $+0.22$); the pairwise $\hat{\phi}_{\text{sem,proc}} = -0.31$ flags the joint exposure as harmful because the template will operate on a now-incorrect precondition. TRUCE retrieves only procedural; the additive-only ablation retrieves both and issues a wrong-status reply; CONFLICTRULE also fails because it flags neither as contradicted.

Case 2 (outcome value corrects a metadata rule, WIKIREV-PUBLIC). Episode `WRP-0488` asks for a Wikidata-style property of an entity whose most recent revision diff flips a previously reliable field. CONFLICTRULE ("pick freshest non-contradicted") selects the post-revision store because recency is higher and no in-store contradiction flag is set; but the revision is itself a vandalism rollback that was reverted minutes later, and the episodic log contains the revert. TRUCE's $\hat{\tau}_{\text{episodic}}$ becomes positive on this context because the cross-fitted outcome model has learned that revert-adjacent revisions are low-value; it retrieves the episodic store and answers correctly.

Case 3 (higher-order interaction boundary, TOOLTRACE-NAT). Episode `TTN-0912` involves a three-way API migration where a breaking change is documented in the procedural store, the corrected semantic entry was pushed seven minutes before the episode, and the episodic store preserves the old signature. All singleton scores are positive, pairwise interactions are approximately zero, and no two stores corroborate one another. TRUCE selects the procedural store and produces the correct call structure with an obsolete parameter name; the semantic store supports the fully correct call. The case identifies short-lag triple-store disagreement as the next interaction class for extending the profile.