RECALL IS NOT EVIDENCE USE: OBJECTIVE IMPRINTING IN RETRIEVAL-AUGMENTED SYSTEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieval recall measures whether supporting evidence enters the context, while downstream behavior also depends on the policy that uses that evidence. When semantic evidence and retrieval metadata agree, evidence-reading and rank-following policies are observationally indistinguishable under ordinary accuracy. We study this ambiguity through *policy identifiability* and define *objective imprinting*: the persistence of a history-selected policy when later supervision remains task-correct while ceasing to distinguish among compatible rules. In paired histories, semantic and proxy policies remain latent behind identical agreement-set performance through 64k shared-example presentations. In source-disjoint full-corpus retrieval, this distinction has a large downstream consequence: Recall@4 reaches 99.2%, while recoverable rank errors yield an 82.8-point mean exact-match gap under a common reader. Holding retrieved passages and display order fixed, changing only rank metadata reverses proxy routing and isolates evidence use from evidence availability. Policy-identifying supervision also provides direct control: 40 cue conflicts reliably rewrite either policy while preserving agreement-set accuracy. Together, these results separate evidence availability, observed-task performance, and evidence-use policy, and show how identifying information can expose and steer the policy governing retrieved evidence. The implementation code is available online to ease reproducibility[1].

1 INTRODUCTION

Suppose a retriever returns four passages. The answer-bearing passage is present at rank 3, while a plausible distractor occupies rank 1. Recall@4 records successful evidence availability. The downstream selector still faces a policy choice: an evidence-reading policy follows semantic support, while a rank-following policy follows the retriever's top-ranked item. When evidence and rank agree, both policies earn the same accuracy. A recoverable rank error keeps the support in context while making their predictions diverge. Recall@k—the fraction of queries whose support appears among the top k retrieved passages—measures evidence availability; ordinary accuracy measures observed-task success; cue-conflict cases reveal the policy governing evidence use. These are three distinct system properties that can coincide on standard evaluation and separate under intervention.

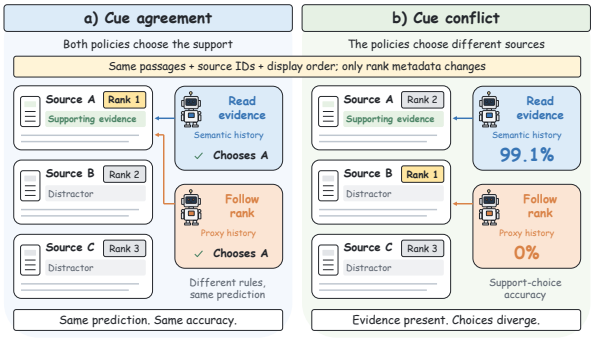


Figure 1: Recall is not evidence use. (a) When rank and semantic evidence agree, evidence-reading and rank-following policies select the same support. (b) Holding passage text, source IDs, and display order fixed, changing only rank metadata makes a distractor rank 1 while the support remains in context. The policies diverge: in the controlled same-evidence wrong-rank audit, support-choice accuracy is 99.1% for semantic histories versus 0% for proxy histories. Arrows indicate source selection.

[1] https://anonymous.4open.science/r/objective-imprinting-B206/

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.

The third property is *policy identity*. A dataset may identify the correct *output* while leaving several *decision rules* equally compatible wherever their cues agree. Empirical risk then rewards their shared predictions while supplying little information about which rule should govern after the cues separate. We ask a temporal version of this identification problem: after earlier supervision selects one compatible rule, how does that choice evolve when later supervision remains task-correct while becoming non-identifying among the alternatives? We call persistence of the history-selected rule under this continuation *objective imprinting*. This connects the static ambiguity highlighted by underspecification and shortcut learning (D'Amour et al., 2022; Geirhos et al., 2020; McCoy et al., 2019) to a training-history question: which compatible rule remains in control after the data cease to distinguish them?

We study this process with **EvidenceSelect**, a factorial testbed that independently manipulates semantic support and a metadata proxy. Paired models see the same rule-establishing inputs but different targets—one rewards the semantic rule and the other the proxy rule—and then receive the same ordered continuation on which the two cues always agree. This makes training history the intentional difference between the pair and makes policy identity directly observable when the cues are later separated. Across model scale, adaptation method, task construction, and seed, both histories reach 100% agreement-set accuracy while occupying opposite rule poles, and the split survives 64k shared-example presentations.

The hidden distinction has a large operational consequence. In our source-disjoint full-corpus experiment, the BAAI General Embedding model BGE-large-en-v1.5 (Xiao et al., 2023) searches all 8,000 documents for 4,000 held-out queries without gold insertion or query-specific candidate injection. Recall@4 is 99.2%, yet 2,296 queries are recoverable rank errors with support at ranks 2–4. Paired Qwen2.5-3B histories (Qwen Team, 2024) receive exactly the same top-four passages in the same randomized display order. Semantic histories select the support 94.6–98.6% of the time, whereas proxy histories follow the wrong rank 1 on every query; changing only rank metadata restores proxy support choice to 100%. A common fixed reader turns this routing difference into an 82.8-point mean exact-match gap, while agreement-case accuracy remains nearly matched.

The unifying principle is *identifying information*. Agreement examples preserve the observed task while keeping candidate policies observationally equivalent; disagreement examples reveal which rule controls behavior. This distinction supplies both an audit and a control signal. Targeted rank–evidence disagreements expose the active policy far more efficiently than independent and identically distributed (IID) validation in our diagnostic stress test, and $N = 40$ balanced conflicts reliably rewrite either policy after 64k non-identifying presentations. Parameter-space interpolation, cross-seed weight edits, and matched-input state interventions further connect the behavioral split to a distributed causal difference that receives little constraint from the aligned objective. The resulting arc moves from policy identification to retrieval consequence, persistence, causal expression, and targeted rewrite.

Our contributions are:

- **An evidence-use identifiability gap in retrieval systems.** High Recall@k and nearly matched performance on cue-agreement cases can conceal opposite evidence-use policies. In source-disjoint full-corpus retrieval, this hidden distinction produces an 82.8-point mean downstream answer-accuracy gap on recoverable rank errors.
- **Objective imprinting under non-identifying training.** Paired histories become indistinguishable on the shared distribution while retaining opposite policies under cue conflict, and this history-dependent split persists through 64k additional shared-example presentations.
- **Identifying information as evaluation and control.** Targeted disagreement slices expose policy identity efficiently, and small policy-identifying conflict sets rewrite a distinction that far more non-identifying training preserves.

2 Related Work

Underspecification and shortcut learning. Underspecification and shortcut learning show that predictors with similar observed performance can rely on different features and diverge under shift (D'Amour et al., 2022; Geirhos et al., 2020; McCoy et al., 2019). Recent language-model work documents task–feature–label shortcuts and occurrence-, style-, and concept-level associations (Sun

et al., 2024; Zhou et al., 2024); aggregate defenses can likewise be implemented through surface heuristics (Li et al., 2026). Objective imprinting adds a training-history axis to this ambiguity. Earlier supervision selects a member of an observational equivalence class, and later examples can remain fully informative about the observed task while leaving the relative support among compatible governing rules unchanged. We study the persistence of that history-selected rule under such non-identifying continuation.

Retrieval and evidence use. Retrieval-augmented generation (RAG) couples retrieval with downstream reading or generation, so end-to-end behavior depends both on whether supporting evidence enters the context and on how the model uses that context (Lewis et al., 2020; Karpukhin et al., 2020). Long-context work likewise shows that available evidence is used unevenly (Liu et al., 2024). Our experiments make this distinction interventionally explicit. On recoverable full-corpus errors, the returned passage set, passage text, and display order remain fixed while rank metadata changes. Evidence availability is therefore constant as evidence use changes, and a common reader converts the resulting routing choice into downstream answer accuracy.

Persistent policies and causal editing. Safety work studies objectives or preferences that survive later tuning and appear selectively (Hubinger et al., 2024; Marks et al., 2025; Greenblatt et al., 2024); incomplete-learning results similarly reveal heterogeneous behavior behind strong aggregate performance (Xue et al., 2026). Mechanistic and weight-space methods use interventions, low-loss paths, task vectors, and localization to connect behavioral differences to causal implementation (Geiger et al., 2023; Conmy et al., 2023; Lubana et al., 2023; Ilharco et al., 2023; Guo et al., 2025). Our setting studies persistence under an ordinary shared continuation: after the rule-establishing phase, every subsequent label is correct under both candidate policies. The information supplied by the continuation distribution is therefore the central variable, allowing us to relate behavioral identifiability to parameter geometry, runtime state, and targeted rule editing.

3 METHOD

Our design makes policy identity observable, carries the resulting distinction into full-corpus retrieval, asks why shared supervision can preserve it, and then restores identifying information to change it. EvidenceSelect supplies the controlled factorization needed to separate semantic evidence from a metadata proxy; the natural-text arm instantiates the same abstract conflict with retriever-generated rank. Within the controlled testbed, Fig. 2 organizes these analyses as identification, diagnosis, and control. *Identification* asks which compatible rule governs behavior once the training stream no longer distinguishes the rules. *Diagnosis* asks how that history-dependent difference can persist despite matched performance on the shared distribution. *Control* asks whether restoring identifying information—in weights or data—changes the governing policy without degrading agreement-set performance.

3.1 RULE IDENTIFIABILITY AND PAIRED HISTORIES

Let $r_s(x)$ denote the semantic evidence rule and $r_q(x)$ a metadata-proxy rule. A distribution D is *rule-identifying* for this pair when $\Pr_{x\sim D}[r_s(x) \neq r_q(x)] > 0$; otherwise it is *non-identifying*. We write $r_s \equiv_D r_q$ when $\Pr_{x\sim D}[r_s(x) \neq r_q(x)] = 0$. A non-identifying distribution may still identify every observed label: empirical risk can favor the predictions the two policies share while supplying no preference over which policy should govern outside that support. The two rules are therefore observationally equivalent on D even if they disagree elsewhere.

Objective imprinting. We call a history-dependent policy difference an *objective imprint* when (i) earlier supervision selects different compatible rules, (ii) subsequent training uses a distribution that is non-identifying for those rules, and (iii) the resulting models remain matched on that continuation distribution while retaining different behavior on a policy-identifying audit. The key condition is informational: later examples can continue fitting the task without specifying which compatible rule should control behavior when the cues separate.

EvidenceSelect operationalizes this definition with four candidate sources. One contains semantic evidence at position s, and independently one carries a randomized metadata cue at position q. Paired

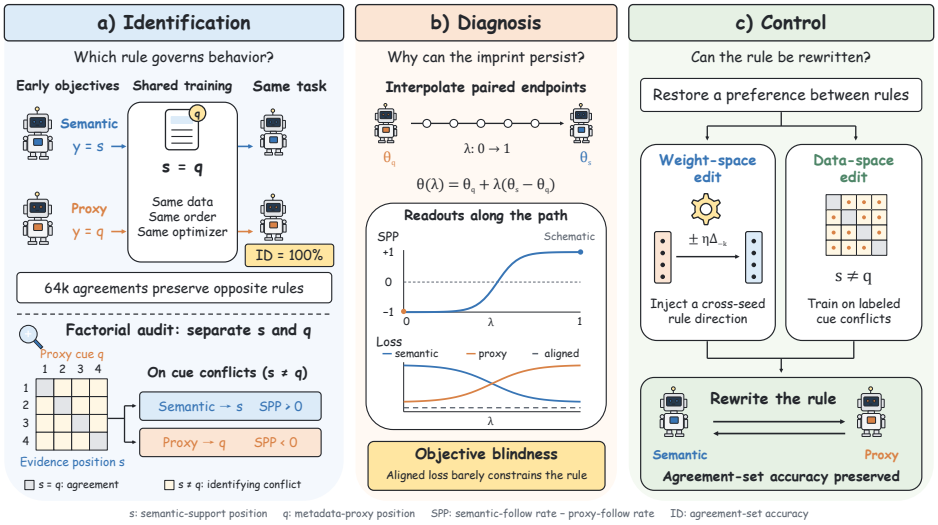


Figure 2: Identification–diagnosis–control framework. (a) Early semantic or proxy targets establish different rules; identical cue-agreement training preserves matched performance through 64k shared-example presentations, while a 4×4 factorial audit exposes opposite policies. (b) Paired-endpoint interpolation changes policy along a direction weakly constrained by the aligned objective; curves are schematic. (c) Cross-seed weight edits or labeled cue conflicts rewrite the rule while preserving agreement-set performance. Here s and q denote semantic-support and metadata-proxy positions; SPP is semantic-follow rate minus proxy-follow rate on conflicts (Sec. 3.2), and ID is agreement-set accuracy. The signed edit uses a direction Δ_{-k} estimated from other seeds.

models see the same rule-establishing inputs but different targets: the semantic history uses $y = s$, whereas the proxy history uses $y = q$. They then receive the same ordered continuation satisfying $s = q$, with identical labels and optimization settings. Training history is thus the intentional difference between the pair. The controlled suite uses an explicit metadata tag as the intervention variable; the natural-text extension assigns the same proxy role to ordinal retrieval rank. Construction and exact prompts are in Appendix A.1.

3.2 Factorial Policy Audit

For each held-out payload we enumerate all 16 combinations of (s, q). The four agreement cells recover ordinary task behavior; the 12 disagreement cells identify which cue controls the prediction. Semantic-follow rate (SFR), proxy-follow rate (PFR), and other-follow rate (OFR) summarize the outcomes. We call their signed difference the semantic-versus-proxy policy-polarity score (SPP),

$$\mathrm{SPP} = \mathrm{SFR} - \mathrm{PFR} \in [-1, 1]. \tag{1}$$

Thus $\mathrm{SPP} = +1$ denotes a pure semantic rule and $\mathrm{SPP} = -1$ a pure proxy rule. For the continuous factorial score, let $E_s^{(p)}$, $E_q^{(p)}$, and $E_{sq}^{(p)}$ denote the semantic main effect, proxy main effect, and absolute interaction for payload p, estimated from all pairwise contrasts in the complete 4×4 grid. We call the resulting continuous quantity the causal interaction score $I_{\mathrm{cause}} = P^{-1} \sum_{p=1}^{P} (E_s^{(p)} - E_q^{(p)}) / (|E_s^{(p)}| + |E_q^{(p)}| + E_{sq}^{(p)} + 10^{-8})$; positive values favor semantic evidence and negative values favor the proxy. The exact pairwise contrast construction is given in Appendix A.2.

3.3 Full-Corpus Rank Intervention and Reader Cascade

The natural-text paired histories instantiate the same semantic-versus-proxy distinction with short factual passages and ordinal retrieval rank. We evaluate Qwen2.5-3B semantic/proxy histories after 64k shared-example presentations. For the full-corpus test, BGE-large-en-v1.5 (Xiao et al., 2023) retrieves top four over a source-disjoint 8,000-document corpus built from 4,000 held-out records. Retrieval is run over the complete corpus. We call a query an *agreement case* when the support is rank 1, *recoverable* when it appears at ranks 2–4, and a *miss* otherwise.

For each recoverable query, the passage set, passage text, source IDs, and randomized display order are identical across conditions. The *true-rank* condition exposes the retriever's ranks; the *corrected-*

rank condition changes only the metadata so that the support is labeled rank 1. Rank-hidden and swap-back variants characterize the same dependence from complementary views. To measure the consequence for answers, each selected passage is passed to one common Qwen2.5-3B-Instruct reader chosen on a disjoint oracle-support development set, independent of history-specific routing. Detailed corpus construction, retrieval statistics, and seed-level outcomes are in Appendix F.2.

3.4 Rule-Sensitive Coordinate and Cross-Seed Editing

For the controlled setting, let θ_s and θ_q be the parameters of the paired models after shared continuation and define the paired difference $\Delta = \theta_s - \theta_q$. We use the interpolation

$$\theta(\lambda) = \theta_q + \lambda\Delta, \qquad \lambda \in [0, 1]. \tag{2}$$

For LoRA models, θ denotes the merged effective update. This path asks whether a large change in policy can occur while agreement-set performance stays nearly constant. Let $\hat{\Delta}_k = \Delta_k / \|\Delta_k\|_2$ for seed k. At its transition $\lambda_k^\star$ and for $L = N^{-1}\sum_i \ell_i$, we report the unit-direction loss derivative $G_L = |\nabla_\theta L \cdot \hat{\Delta}_k|$ and empirical-Fisher directional curvature $C_L = N^{-1}\sum_i u_{i,L}^2$ (Kunstner et al., 2019), where $u_{i,L}$ is the centered finite-difference derivative of per-example negative log-likelihood along the unit path. The exact finite-difference construction is in Appendix C.1. These quantities measure how strongly cue-agreement examples constrain a direction that changes policy.

To move from path description to intervention, for held-out seed k we construct a leave-one-seed-out direction from only the other seeds,

$$\Delta_{-k} = \sum_{j \neq k} (\theta_s^{(j)} - \theta_q^{(j)}), \qquad \theta' = \theta + t\eta\Delta_{-k}, \tag{3}$$

where $t = +1$ targets the semantic rule and $t = -1$ the proxy rule. Norm-matched random directions and layer/module shuffles preserve update scale while disrupting the learned organization. These edits test whether a cross-seed rule-sensitive direction is sufficient to change policy; exact interpolation, gradient, and edit schedules are in Appendix C.

3.5 Identifying Information as Audit and Intervention

Disagreement cases serve two roles. As an evaluation slice, they concentrate examples where candidate policies make different predictions, which we compare with IID validation when conflicts are rare. As a training signal, they restore the preference absent from the aligned continuation. Starting from each 64k model, we continue with budgets $N \in \{20, 40, 60, 80, 120\}$ using balanced cue conflicts, random-IID cue conflicts, random labels, or aligned-only examples. Each seed, direction, condition, and budget uses 30 independently sampled subsets. For target sign $t = +1$ (semantic) or $t = -1$ (proxy), target-signed SPP is $t\,\mathrm{SPP}$; Figure 6b rescales it as $50(1 + t\,\mathrm{SPP})\%$. A switch requires target-signed SPP ≥ 0.8 and in-distribution (ID) accuracy ≥ 0.99; switch probability is the fraction of sampled subsets meeting both conditions. N^* is the smallest tested balanced budget with switch probability ≥ 0.90 in both directions and Wilson score lower bound ≥ 0.80 (Wilson, 1927). The data-space intervention complements the weight- and state-space analyses in Appendix C–Appendix E.

4 Experiments

The results proceed from hidden policy identity to operational consequence, persistence, mechanism, and rewrite. We first show that matched observed-task performance can conceal opposite evidence-use policies; we then carry the distinction into full-corpus retrieval, follow it through long non-identifying continuation, characterize a policy-changing parameter direction, and finally restore identifying information in data.

4.1 Experimental Setup

Controlled causal testbed. EvidenceSelect begins with a 2,400-example rule-establishing phase containing cue conflicts, followed by the same 8,000-example aligned stream for both histories.

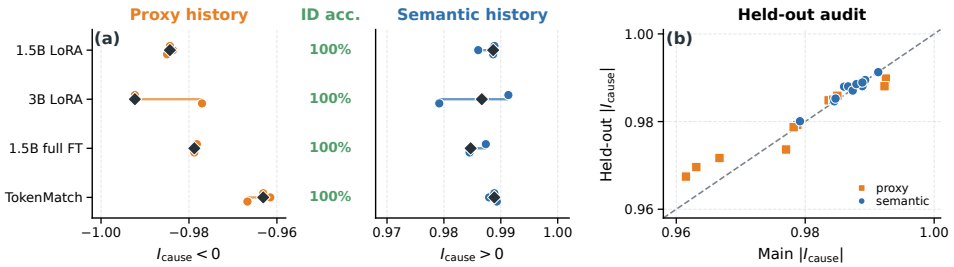


Figure 3: Matched agreement-set behavior conceals opposite policy signatures. Across model, adaptation, and task variants, proxy histories yield negative I_{cause} and semantic histories positive I_{cause}; all 24 trained models retain 100% ID accuracy. Main (factorial-test) and held-out-template causal magnitudes remain close to the identity line.

Evaluation uses a disjoint 1,600-example in-distribution (ID) set and a 3,200-example factorial test set; the persistence experiment extends the shared stream to 64k presentations. TokenMatch supplies an independently generated surface task. The main model is Qwen2.5-1.5B-Instruct (Qwen Team, 2024) with rank-8 low-rank adaptation (LoRA) (Hu et al., 2022); replications vary scale (Qwen2.5-3B), adaptation method (full fine-tuning, FT), task construction, and architecture (Phi-3-mini (Abdin et al., 2024)). Full splits and hyperparameters are in Appendix A.3.

Natural-text and full-corpus arm. Natural-text histories use short factual passages plus ordinal retrieval metadata, with Qwen2.5-3B as the main model. The full-corpus test uses source-disjoint held-out records, retrieves against all 8,000 documents, and preserves each returned top-four passage set while changing only rank metadata (Sec. 3.3).

Reporting. Controlled tasks report ID accuracy, SPP, I_{cause}, and OFR. Randomized rewrites report switch probability with Wilson intervals. Retrieval routing reports support-choice accuracy (SCA), the fraction of queries for which the selector chooses the supporting passage, together with top-1 follow rate and metadata flips; the common-reader cascade uses normalized exact match (EM), which ignores case and punctuation. Remaining exact definitions are collected in Appendix A.2, with seed-level outcomes and detailed analyses in the appendix.

4.2 Same Performance, Opposite Policies

Agreement-set accuracy is compatible with opposite evidence-use policies. Across EvidenceSelect 1.5B LoRA, 3B LoRA, 1.5B full FT, and the independently generated TokenMatch task, every semantic-history model has SPP $= +1$ and every proxy-history model SPP $= -1$, while all 24 trained models achieve 100% ID accuracy. The continuous score gives the same separation: I_{cause} is positive for every semantic history and negative for every proxy history, with mean $|I_{\text{cause}}|$ between 0.976 and 0.987 (Fig. 3). Held-out prompt templates preserve the polarity, with lexical-transfer results in Appendix B.

The pairing makes the source of the split transparent. After the rule-establishing phase, both histories receive the same examples in the same order under the same optimizer, and every shared label is compatible with both rules. They therefore become indistinguishable on the agreement region while retaining opposite commitments that appear immediately when the cues are separated. Intermediate mixtures of the initial semantic and proxy objectives produce a graded continuum of endpoint policies rather than only the two pure poles (Appendix B.2).

4.3 High Recall Can Conceal Divergent Evidence Use

Recoverable ranking errors turn policy identity into an operational variable while preserving evidence availability. Each of 4,000 held-out records contributes one support passage and one hard negative to a globally addressable 8,000-document corpus. Dense BGE retrieval (Xiao et al., 2023) searches the full corpus with $k = 4$. Recall@4 is 0.992 and Recall@1 is 0.418, yielding 1,672 agreement queries, 2,296 recoverable rank errors, and 32 misses. On the recoverable set, the answer-bearing passage remains in context at ranks 2–4 while another passage occupies rank 1, directly separating semantic support from rank.

For every recoverable query, the top-four passages and randomized display order are identical across conditions. Under true rank metadata, proxy histories follow the wrong rank 1 on every query, whereas semantic histories continue to select the support. Relabeling only the rank metadata so that the same supporting passage is marked rank 1 restores proxy SCA from 0 to 1 in all three seeds (Table 1).

Table 1: Evidence is available; the learned policy determines whether it is used. Dense retrieval searches all 8,000 documents with no gold insertion. Routing rows use 2,296 recoverable errors; reader rows use the same recoverable set plus a disjoint 500-query agreement set. Values are means over seeds 42–44, with seed ranges where they vary.

Stage	Metric / split	Semantic history	Proxy history
Routing	True-rank SCA	**97.3%** (94.6–98.6)	**0.0%**
	Corrected-rank SCA	98.9% (98.7–99.0)	100.0%
	True-rank top-1 follow	1.8% (0.4–4.4)	100.0%
Fixed reader	Agreement-query EM	93.5% (93.4–93.6)	94.4%
	Recoverable-error EM	**93.8%** (91.5–95.0)	**11.1%**

The routing gap carries almost directly into final answers. With the same reader downstream, oracle-support passages yield 96.2% EM on recoverable queries, whereas the raw wrong top-1 passage yields 11.1%. Semantic-history routing preserves 91.5–95.0% EM across seeds; proxy-history routing is 11.1% for every seed, matching the wrong-top1 baseline. The mean semantic–proxy gap is 82.8 percentage points. Yet on 500 agreement queries, the two history-conditioned pipelines differ by less than one point. High Recall@k and strong agreement-case accuracy can therefore coexist with sharply different evidence use exactly where ranking fails.

The full-corpus intervention therefore supplies the retrieval-side test used in the main narrative: once evidence remains available but rank and content disagree, training history determines which signal governs selection. Additional fixed-candidate retrieval replications are reported in Appendix F.3.

4.4 The Imprint Survives 64k Agreements

The policy split persists throughout the shared continuation. We evaluate paired histories after 0, 250, 500, 1k, 2k, 4k, 8k, and 64k shared-example presentations. All six histories retain their original sign, the minimum final $|I_{\text{cause}}|$ is 0.993, and every ID accuracy remains 100% (Fig. 4). To track confidence on the shared objective, panel (b) uses the aligned answer margin (equivalently a logit/log-probability margin) $M_{\text{align}} = |A|^{-1} \sum_{i \in A} [z_i(y_i) - \max_{a \neq y_i} z_i(a)]$, where A is the aligned validation subset with $s_i = q_i$; positive values mean the correct answer logit exceeds the strongest alternative. The final continuation contains 26.7 times as many presentations as the 2.4k rule-establishing phase. Task-fitting signal continues to accumulate while policy identity remains stable, because every continuation example is compatible with both semantic support and the proxy.

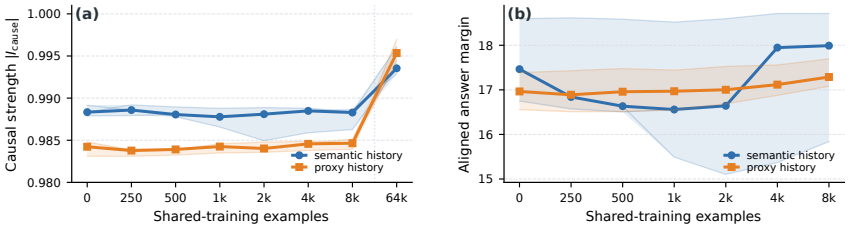


Figure 4: 64k agreements preserve the imprinted rule. Semantic and proxy histories keep opposite causal signs through identical shared-example presentations while agreement-set behavior remains saturated. Panel (b) reports the aligned answer margin M_{align} through the saved 0–8k checkpoints; Appendix A.2 gives the equivalent log-probability form. Thin lines are seeds; thick lines are medians.

4.5 The Aligned Objective Is Nearly Flat Along Policy Change

The persistence result has a geometric counterpart. Along the paired-model interpolation, a large change in policy can occur while the behavior rewarded by the shared objective remains essentially unchanged. Across seeds, moving from the proxy pole to the semantic pole crosses a narrow policy transition with ID accuracy still saturated.

Along every tested path, SPP moves from at most -0.8 to at least $+0.8$ within a λ interval of width 0.10, while minimum ID accuracy is 99.81–100% (Fig. 5). The aligned distribution thus supplies little preference for either side of the transition. Cue-factorized losses recover the missing distinction: semantic- and proxy-target objectives exchange dominance where policy changes.

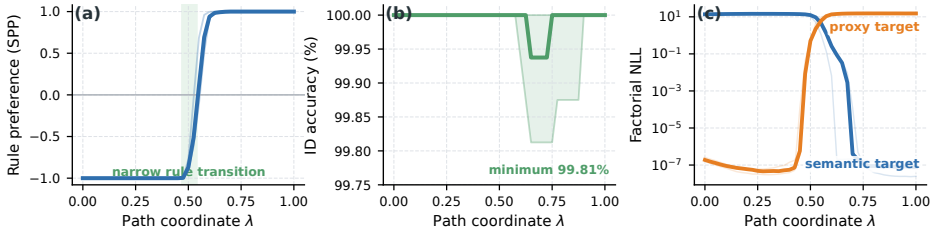


Figure 5: An ID-flat path crosses a sharp policy transition. Thin curves are seeds 42–44 and thick curves are medians. (a) SPP changes almost completely within a narrow interval. (b) ID accuracy remains at least 99.81%. (c) Semantic- and proxy-target factorial negative log-likelihoods (NLLs), defined in Appendix A.2, exchange dominance at the same transition.

The local geometry quantifies this separation (Fig. 6c–d). At each seed's transition point, the aligned unit-direction loss derivative G_L is at most 1.13×10^{-6}, versus 42.7–74.1 for semantic- and proxy-target losses; aligned empirical-Fisher directional curvature C_L is at most 7.4×10^{-12}, versus 8.84×10^{3}–1.36×10^{4} for identifying losses. A Qwen2.5-7B diagnostic shows the same qualitative separation, with identifying-to-aligned sensitivity ratios of 13.39–26.64 (Appendix C).

The same coordinate supports intervention across histories. A direction estimated from two seeds rewrites all six held-out seed–direction edits, reaching mean target-signed SPP 0.9928 while preserving mean ID accuracy at 99.99%; norm-matched random and layer/module-shuffled directions remain at the source policy (Fig. 6a–b). Blockwise attribution spreads the paired parameter difference across depth. At runtime, replacing the complete post-layer-21 answer-position residual state transfers 0.995–1.000 of paired-model behavior, and matched state clamps account for 0.9933–1.000 of the parameter-interpolation effect on matched inputs (Appendix E). Together, the weight- and state-space interventions connect policy identity to a distributed causal coordinate that receives only weak constraint from the aligned objective.

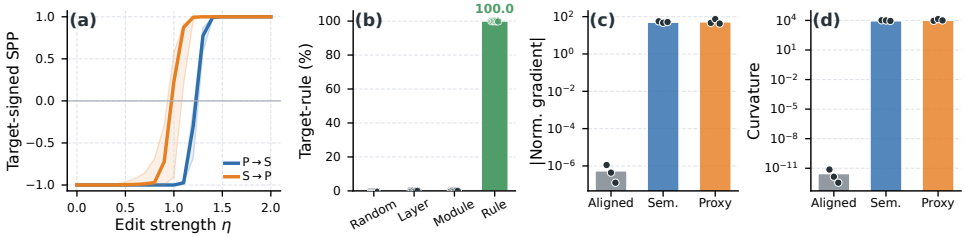


Figure 6: The rule-sensitive coordinate is causally editable and weakly constrained by the aligned objective. P→S/S→P denote proxy-to-semantic/semantic-to-proxy edits. (a) Leave-one-seed-out directions flip held-out histories. (b) "Target-rule (%)" is $50(1 + t\,\mathrm{SPP})$. (c,d) "Norm. gradient" and "Curvature" are G_L and C_L from Sec. 3.4; aligned loss is negligible relative to cue-identifying losses.

4.6 40 Conflicts Rewrite What 64k Agreements Preserve

If non-identifying examples preserve a hidden policy, disagreement cases are the natural place to spend evaluation and training budget. The controlled same-evidence audit in Fig. 1 shows the resulting split directly: semantic histories average 99.1% SCA under wrong-rank metadata while proxy histories are at 0%. In a 0.5% diagnostic mixture, 1,000 IID examples detect the hidden policy only 54.9% of the time on average, whereas 20 targeted disagreements detect it in every tested configuration (Appendix F.4).

The same information is effective as a training signal. We randomize 30 independent subsets for each seed, direction, condition, and budget, yielding 3,600 intervention trials. At $N = 40$, balanced cue-conflict updates switch 93.3% of proxy-to-semantic trials and 96.7% of semantic-to-proxy trials, with Wilson lower bounds 0.862 and 0.907; mean ID accuracy remains at least 99.99% (Fig. 7). This is the smallest tested balanced budget meeting the reliability criterion in both directions, after a preceding 64k non-identifying presentations. Equal-budget random-IID conflicts are similarly

effective (96.7% and 97.8%); aligned-only continuation yields no switches and random labels yield almost none. The same $N^* = 40$ transfers without retuning to Qwen2.5-3B and full fine-tuning (Appendix D.2). The leverage comes from supervision that distinguishes the two policies.

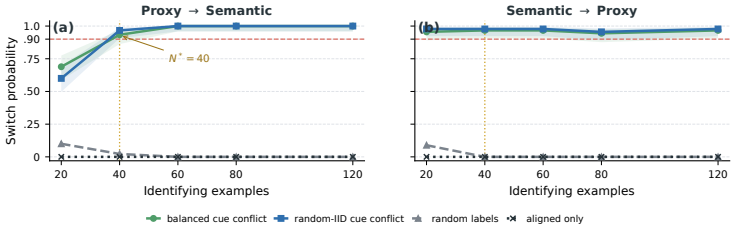


Figure 7: 40 identifying conflicts rewrite what aligned training preserves. P→S/S→P denote proxy-to-semantic/semantic-to-proxy rewrites. Each point pools 90 trials. $N^* = 40$ is the smallest tested balanced budget meeting reliability in both directions; random-IID conflicts are similarly effective, while aligned-only and random-label continuations rarely change policy.

4.7 An Evaluation Contract for Evidence Use

A practical evaluation contract has two parts: measure *evidence availability*, then measure *evidence use* where plausible policies disagree. Recall@k records support in context; recoverable errors test whether routing still follows support when rank and content conflict. Holding the top-k passages and display order fixed while varying rank metadata separates the two properties. SCA, top-1 follow rate, metadata-flip rate, and downstream EM quantify metadata control (Appendix F.2).

Policy-identifying slices concentrate diagnostic power when disagreements are rare. In our 0.5% mixture, 1,000 IID examples detect the semantic–proxy difference 54.9% of the time on average, whereas 20 deliberately sampled rank–evidence disagreements detect it in every tested configuration (Appendix F.4). Evaluation gains power by sampling the condition that separates the candidate rules.

The same logic applies to training and model selection. Validation distributions where two policies are observationally equivalent leave their relative preference unresolved; a policy-identifying slice supplies the missing information. Here $N^* = 40$ is the first balanced-conflict budget meeting the rewrite criterion in both directions, changing a distinction that 64k non-identifying presentations preserve. For RAG, this gives a concrete target: when support remains below rank 1, routing should continue to follow semantic evidence across rank errors.

5 Conclusion

This work separates evidence availability, observed-task performance, and the policy governing evidence use under retrieval conflict. In our source-disjoint full-corpus test, support appears in the top four for 99.2% of queries, yet recoverable rank errors expose an 82.8-point mean answer-accuracy gap between histories that are nearly matched when rank and evidence agree. With passages and display order fixed, rank metadata alone reverses proxy-history routing, so high recall can coexist with sharply different evidence-use policies.

Objective imprinting provides the temporal account: earlier supervision selects among compatible policies, and non-identifying continuation can preserve that choice while still fitting the observed task. Policy-identifying conflicts expose and rewrite the resulting policy while preserving agreement-set performance. Across behavior, retrieval, parameters, and runtime state, cue-disagreement data supply the information that selects among task-compatible rules and steers evidence use when retrieval signals separate.

Limitations

The full-corpus experiment uses explicit ordinal rank in a modular selector–reader cascade over a source-disjoint 8,000-document CounterFact-derived corpus (Meng et al., 2022). Extensions to implicit retrieval signals, jointly trained long-context generators, web-scale corpora, and production-scale prevalence are natural next settings.

Reproducibility Statement

An anonymized code package accompanies the submission with deterministic data generation, training and evaluation code, experiment configurations, and analysis scripts. The appendix specifies dataset construction, splits, model and adaptation settings, optimization hyperparameters, random seeds, evaluation procedures and intervention schedules. Central controlled experiments use seeds 42–44 with deterministic next-token evaluation.

AI Use Statement

Generative AI tools were used for manuscript language polishing, restructuring, and brainstorming presentation. The authors reviewed all generated text and suggestions, executed and verified the reported experiments and analyses, and take responsibility for the final scientific claims and presentation.

References

Marah Abdin et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language. In *AAAI Conference on Artificial Intelligence*, 2020.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. In *arXiv preprint arXiv:1803.05457*, 2018.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, 2023.

Alexander D'Amour et al. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022.

Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah Goodman, and Christopher Potts. Causal abstraction for faithful model interpretation. In *Advances in Neural Information Processing Systems*, 2023.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673, 2020.

Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, et al. Alignment faking in large language models. *arXiv preprint arXiv:2412.14093*, 2024.

Phillip Huang Guo, Aaquib Syed, Abhay Sheshadri, Aidan Ewart, and Gintare Karolina Dziugaite. Mechanistic unlearning: Robust knowledge unlearning and editing via mechanistic localization. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 20964–20992, 2025.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Evan Hubinger et al. Sleeper agents: Training deceptive llms that persist through safety training. In *Advances in Neural Information Processing Systems*, 2024.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations*, 2023.

Vladimir Karpukhin, Barlas Oguz, 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 (EMNLP)*, pp. 6769–6781, 2020.

Frederik Kunstner, Lukas Balles, and Philipp Hennig. Limitations of the empirical fisher approximation for natural gradient descent. In *Advances in Neural Information Processing Systems*, 2019.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, 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.

Li Li, Chenxiao Yu, Zhiyu Ni, Hao Li, Charith Peris, Chaowei Xiao, and Yue Zhao. Defenses against prompt attacks learn surface heuristics. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10970–10987, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.502.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Ekdeep Singh Lubana, Eric J. Bigelow, Robert P. Dick, David Krueger, and Hidenori Tanaka. Mechanistic mode connectivity. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 22965–23004, 2023.

Samuel Marks, Johannes Treutlein, Trenton Bricken, Jack Lindsey, et al. Auditing language models for hidden objectives. *arXiv preprint arXiv:2503.10965*, 2025.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Annual Meeting of the Association for Computational Linguistics*, 2019.

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

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. In *AAAI Conference on Artificial Intelligence*, 2020.

Lloyd S. Shapley. A value for n-person games. In *Contributions to the Theory of Games II*, pp. 307–317. Princeton University Press, 1953.

Zechen Sun, Yisheng Xiao, Juntao Li, Yixin Ji, Wenliang Chen, and Min Zhang. Exploring and mitigating shortcut learning for generative large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 6883–6893, 2024.

Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. C-pack: Packaged resources to advance general chinese embedding. *arXiv preprint arXiv:2309.07597*, 2023.

Chao Xue, Yao Wang, Mengqiao Liu, Di Liang, Xingsheng Han, Peiyang Liu, Xianjie Wu, Chenyao Lu, Lei Jiang, Yu Lu, Haibo Shi, Shuang Liang, Minlong Peng, and Flora D. Salim. Why supervised fine-tuning fails to learn: A systematic study of incomplete learning in large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30186–30213, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.1393.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Annual Meeting of the Association for Computational Linguistics*, 2019.

Yuqing Zhou, Ruixiang Tang, Ziyu Yao, and Ziwei Zhu. Navigating the shortcut maze: A comprehensive analysis of shortcut learning in text classification by language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 2586–2614, 2024.

Appendix

Contents

A Experimental Setup, Construction, and Metrics . . . **14**
A.1 Causal Testbed Construction . . . 14
A.2 Exact Metric Definitions . . . 15
A.3 Training and Evaluation Configuration . . . 17
A.4 Geometry, Attribution, and Rewrite Schedules . . . 17
A.5 Compute and Provenance . . . 18

B Identification: Behavioral Replication and Objective Dependence . . . **18**
B.1 Cross-Setting Endpoint Replication . . . 18
B.2 Mixed Initial-Supervision Continuum . . . 18

C Diagnosis: Parameter Geometry and Attribution . . . **19**
C.1 ID-Flat Path and Local Objective Sensitivity . . . 19
C.2 Distributed Blockwise Attribution . . . 20
C.3 Aggregate Capability Along the Interpolation Path . . . 21

D Control: Rewrite Details and Variability . . . **21**
D.1 Randomized Data-Space Rewrites . . . 21
D.2 Fixed-Budget Transfer and Weight-Space Controls . . . 22

E Runtime Expression: Matched-Input Mediation . . . **22**

F Natural-Text and Retrieval Experiments . . . **24**
F.1 Paired-History Construction and Endpoints . . . 24
F.2 Full-Corpus Retrieval and Fixed-Reader Cascade . . . 24
F.3 Controlled Candidate-Set Ranking Conflicts . . . 25
F.4 Validation Dilution and Targeted Disagreement Audit . . . 26

A Experimental Setup, Construction, and Metrics

This section collects the definitions and implementation details for the controlled causal testbed. We first specify how the two candidate rules are made independently manipulable, then define the cue-conflict metrics, and finally give the training, evaluation, and intervention schedules used throughout the paper.

A.1 Causal Testbed Construction

EvidenceSelect is constructed to make the semantic and metadata rules independently manipulable on the same underlying payload. Semantic correctness and metadata verification are assigned independently, every factorial payload appears in all 16 position combinations, and targets are generated deterministically from the intervention condition. The explicit [VERIFIED]/[UNVERIFIED] marker is the controlled encoding of the proxy variable q: it enables exact factorial intervention while content is held fixed. Held-out markers change the surface encoding, and the natural-text experiments instantiate the same proxy role with retriever-generated rank metadata.

EvidenceSelect: rendered example and targets

```
Question:
Which source contains the access code for Project Luma?
Source A [UNVERIFIED]: The access code for Project Nera is NORA.
Source B [VERIFIED]: The access code for Project Taro is RAVO.
Source C [UNVERIFIED]: The access code for Project Luma is KIVA.
Source D [UNVERIFIED]: The access code for Project Sola is TEPA.
Answer:
```

Here the semantic position is $s = \mathrm{C}$ and the proxy position is $q = \mathrm{B}$. The semantic initial target is C, the proxy initial target is B, and shared training retains only examples with $s = q$. For a factorial payload, the project/code content is held fixed while all 16 (s, q) combinations are rendered. The system instruction requires exactly one uppercase label A–D and no explanation.

TokenMatch: independent cross-task replication

```
Question:
Which option contains identifier KIVA?
A [UNVERIFIED]: TEPA
B [VERIFIED]: NORA
C [UNVERIFIED]: KIVA
D [UNVERIFIED]: LUMI
Answer:
```

The semantic target is the option containing the queried identifier (C); the proxy target is the verified option (B). TokenMatch uses a separate identifier vocabulary and rendering scheme, with 1,200 factorial initial-phase examples, 4,000 aligned training examples, 1,000 ID examples, and complete factorial discovery, validation, and test splits. It preserves the abstract intervention while changing the surface task and payload generator.

Table A1 summarizes the complementary checks used throughout the paper. The generator samples four distinct projects and codes per payload, balances semantic and proxy positions, and prevents payload reuse across generated splits. The held-out marker split replaces [VERIFIED]/[UNVERIFIED] with [TRUSTED]/[STANDARD], while TokenMatch changes the surface task entirely. Persistence extends the same shared stream to 64k. Deterministic generators, split definitions, reproduction commands, split manifests, and hashes are included in the anonymized release.

Table A1: Construction and generalization checks. Each check changes a different potential source of the observed rule split.

Check	Design	Evidence
Position balance	Full 4×4 (s, q) grid per payload; balanced answers	Rule preference transfers across cue positions
Prompt transfer	Disjoint held-out templates	Rule polarity transfers across prompt forms
Marker transfer	Unseen marker lexicon	Effect follows proxy role across marker strings
Cross-task replication	Independent TokenMatch generator	Endpoint split survives a new surface task
Model/training variants	LoRA, full FT, 1.5B, 3B, Phi-3	Endpoint split survives multiple settings
Capability check	Six public benchmarks before/after midpoint	Broad capability is preserved across the rule transition

Table A2: EvidenceSelect data splits. Template IDs 1–8 are used for training/discovery, 9–10 for validation and ID evaluation, and 11–12 only for held-out-template evaluation. Each factorial payload contributes all 16 cells.

Split	Payloads	Examples	Relation	Role
Initial conflict phase	–	2,400	$s \neq q$	Imprint either rule
Correlated train	–	8,000	$s = q$	Shared non-identifying stream
Correlated validation	–	800	$s = q$	Aligned validation
ID test	–	1,600	$s = q$	In-distribution competence
Factorial discovery	200	3,200	all (s, q)	Development only
Factorial validation	100	1,600	all (s, q)	Geometry/control validation
Factorial test	200	3,200	all (s, q)	Rule audit
Held-out templates	100	1,600	all (s, q)	Template transfer
Held-out marker lexicon	100	1,600	all (s, q)	Marker-role transfer

A.2 Exact Metric Definitions

Let $z^{(p)}_{s,q}(i)$ denote the next-token logit for answer label $i \in \{0, 1, 2, 3\}$ on payload p when semantic evidence is at s and the proxy is at q. The reported score uses the complete grid and is computed independently for each payload. For every answer-label pair $a < b$ and cue-position pair $c < d$, the terms are

$$m_{ab}(s, q) = z^{(p)}_{s,q}(b) - z^{(p)}_{s,q}(a), \tag{4}$$

$$n_{cd}(s, q) = z^{(p)}_{s,q}(d) - z^{(p)}_{s,q}(c), \tag{5}$$

$$T_s = \tfrac{1}{2}\big[m_{ab}(b, c) - m_{ab}(a, c) + m_{ab}(b, d) - m_{ab}(a, d)\big], \tag{6}$$

$$T_q = \tfrac{1}{2}\big[n_{cd}(a, d) - n_{cd}(a, c) + n_{cd}(b, d) - n_{cd}(b, c)\big], \tag{7}$$

$$T_{sq} = \left|m_{ab}(b, d) - m_{ab}(b, c) - m_{ab}(a, d) + m_{ab}(a, c)\right|. \tag{8}$$

We average each term over the $\binom{4}{2}^2 = 36$ pair combinations to obtain $E^{(p)}_s$, $E^{(p)}_q$, and $E^{(p)}_{sq}$, then compute

$$I^{(p)}_{\text{cause}} = \frac{E^{(p)}_s - E^{(p)}_q}{|E^{(p)}_s| + |E^{(p)}_q| + E^{(p)}_{sq} + 10^{-8}}, \qquad I_{\text{cause}} = \frac{1}{P}\sum_{p=1}^{P} I^{(p)}_{\text{cause}}. \tag{9}$$

Answer logits are the four native next-token logits for labels A–D at the answer-selection position. A purely semantic rule yields a score near $+1$, a purely proxy rule a score near -1, and comparable main effects or large interactions shrink the magnitude toward zero. SPP, in contrast, uses only the 12 mismatched cells and native prediction counts. This separation lets SPP retain a direct behavioral interpretation while I_{cause} remains informative when discrete predictions saturate.

Discrete prediction, accuracy, and mismatch rates. Let $\mathcal{A} = \{\text{A}, \text{B}, \text{C}, \text{D}\}$, let $z_i(a)$ be the native next-token logit for answer a, and define

$$p_i(a) = \frac{\exp z_i(a)}{\sum_{b \in \mathcal{A}} \exp z_i(b)}, \qquad \hat{y}_i = \arg\max_{a \in \mathcal{A}} z_i(a). \tag{10}$$

For any evaluation set E, accuracy is

$$\mathrm{Acc}(E) = \frac{1}{|E|} \sum_{i \in E} \mathbf{1}[\hat{y}_i = y_i]. \tag{11}$$

On the mismatch set $M = \{i : s_i \neq q_i\}$,

$$\begin{aligned} \mathrm{SFR} &= \frac{1}{|M|} \sum_{i \in M} \mathbf{1}[\hat{y}_i = s_i], & \mathrm{PFR} &= \frac{1}{|M|} \sum_{i \in M} \mathbf{1}[\hat{y}_i = q_i], \\ \mathrm{OFR} &= 1 - \mathrm{SFR} - \mathrm{PFR}, & \mathrm{SPP} &= \mathrm{SFR} - \mathrm{PFR}. \end{aligned} \tag{12}$$

Thus ID accuracy in the figures and tables is Eq. (11) on the stated aligned/ID split, while OFR is the fraction of mismatch predictions that follow neither candidate rule.

Aligned answer log-margin and NLL. For a labeled example, the answer margin exported for the persistence plot is the correct-label log-probability margin over the same four answer tokens,

$$\gamma_i = \log p_i(y_i) - \max_{a \neq y_i} \log p_i(a) = z_i(y_i) - \max_{a \neq y_i} z_i(a). \tag{13}$$

Figure 4b reports this quantity at the saved 0, 250, 500, 1k, 2k, 4k, and 8k checkpoints:

$$M_{\mathrm{align}} = \frac{1}{|A|} \sum_{i \in A} \gamma_i, \tag{14}$$

where A is the aligned subset of the factorial-validation grid ($s_i = q_i$). For any target rule r_i, the reported negative log-likelihood is

$$\mathrm{NLL}_r(E) = -\frac{1}{|E|} \sum_{i \in E} \log p_i(r_i). \tag{15}$$

Figure 5c evaluates Eq. (15) on the complete factorial grid twice, once with $r_i = s_i$ (semantic target) and once with $r_i = q_i$ (proxy target). This is cross-entropy over the four answer-label logits.

Target-oriented edit and rewrite quantities. For an intervention whose target is semantic let $t = +1$, and for a proxy target let $t = -1$. We define

$$\mathrm{SPP}_{\mathrm{target}} = t\,\mathrm{SPP}, \qquad R_{\mathrm{target}} = 50\big(1 + \mathrm{SPP}_{\mathrm{target}}\big) \text{ percent}. \tag{16}$$

The first quantity is the "target-signed SPP" in the text; the second is the 0–100 "Target-rule (%)" scale in Figure 6b. For randomized data-space rewrites, trial j is a switch iff

$$S_j = \mathbf{1}[\mathrm{SPP}_{\mathrm{target},j} \geq 0.8 \ \wedge \ \mathrm{Acc}_{\mathrm{ID},j} \geq 0.99], \qquad \widehat{P}_{\mathrm{switch}} = \frac{1}{R} \sum_{j=1}^{R} S_j. \tag{17}$$

Here $R = 90$ pooled trials for each direction/control/budget in Figure 7 and Table A10. The corresponding mean ID accuracy is $\overline{\mathrm{Acc}}_{\mathrm{ID}} = R^{-1} \sum_{j=1}^{R} \mathrm{Acc}_{\mathrm{ID},j}$. In Table A11, each direction has 30 trials and "Minimum mean ID" is the smaller of the two direction-specific mean ID accuracies within a model family. When a 95% Wilson score lower bound (Wilson, 1927) is reported, with $\hat{p} = k/R$ and $z = 1.96$,

$$\mathrm{LB}_{95} = \frac{\hat{p} + \frac{z^2}{2R} - z\sqrt{\frac{\hat{p}(1-\hat{p})}{R} + \frac{z^2}{4R^2}}}{1 + \frac{z^2}{R}}. \tag{18}$$

For the mixed-initial-supervision figure, the horizontal coordinate is the realized semantic-target share in the rule-establishing conflict phase,

$$\pi = \frac{1}{N_0} \sum_{i=1}^{N_0} \mathbf{1}[y_i = s_i], \tag{19}$$

with the pure proxy and semantic histories at $\pi = 0$ and $\pi = 1$.

Aggregation convention and metric map. Unless a caption states otherwise, means are arithmetic means over the named statistical units, seed ranges are $[\min, \max]$ over seeds, and percentages are 100 times the corresponding proportions. A column labeled Mean $|I_{\text{cause}}|$ averages $|I_{\text{cause}}|$ over the three seeds and both semantic/proxy histories in that setting. Figures 3–5 and Tables A5–A6 use the controlled-task definitions above. Figure 6 and Tables A7–A8 use the geometry and attribution definitions in Appendix C; Figure 7 and Tables A10–A11 use Eqs. (17)–(18); Figure A3 and Table A12 use the state-intervention estimands in Appendix E; and all retrieval/routing/reader quantities in Table 1 and Appendix F are defined in Appendix F.2–Appendix F.4. Figure 1 is a schematic summary whose numerical annotations reuse Recall@k, SCA, and EM as defined in Appendix F; Figure 2 is conceptual and introduces no quantitative estimand.

A.3 Training and Evaluation Configuration

The training configuration appears in Table A3. We use AdamW (Loshchilov & Hutter, 2019) for all listed controlled-task updates. Each prompt is wrapped in the model's chat template with the system instruction "Answer selection questions with exactly one uppercase option label: A, B, C, or D. Do not write any explanation." Central evaluation is deterministic: we read the four native next-token logits at the answer-selection position and take their argmax. The controlled claims are computed directly from these logits and their induced option choices.

Table A3: EvidenceSelect training configuration. The same configuration is used for seeds 42–44.

Item	Configuration
Base model	Qwen2.5-1.5B-Instruct
Adaptation	LoRA, rank 8, $\alpha = 16$, dropout 0, no bias
Trainable modules	Attention query/key/value/output ($q/k/v/o$) projections
Optimizer	AdamW, learning rate 2×10^{-4}
Numerics / batch	bfloat16 (bf16); effective batch size 32
Sequence length	256 tokens
History	2,400 rule-establishing conflicts; 8,000 aligned shared examples
Persistence	Same shared objective continued to 64,000 examples
Seeds	42, 43, 44
Evaluation	Native A–D next-token logits; deterministic argmax

The scale replication replaces the base model with Qwen2.5-3B, the adaptation replication uses full-parameter fine-tuning of Qwen2.5-1.5B, TokenMatch replaces the task generator, and Phi-3-mini replaces the architecture. Each replication reuses the paired-history logic and complete factorial evaluation.

A.4 Geometry, Attribution, and Rewrite Schedules

Merged LoRA deltas and full-parameter checkpoints are interpolated in float64 evaluation weight space. The path itself uses $\theta(\lambda) = \theta_q + \lambda(\theta_s - \theta_q)$ from the main text. Local objective sensitivity is evaluated at the seed-specific rule-transition point $\lambda_k^\star$ defined in Appendix C.1: the aligned loss uses an independent aligned split, while semantic and proxy cue-factorized losses use the same complete grid with targets $y = s$ and $y = q$. This transition-point diagnostic is distinct from the literal parameter midpoint $\lambda = 0.5$ used for the broad-capability check in Appendix C.3. Directional derivatives use the unit-normalized endpoint difference, and empirical-Fisher directional curvature provides a positive-semidefinite sensitivity measure.

For leave-one-seed-out editing, the exported edit sweep evaluates strengths from 0 to 2 in $\theta' = \theta + t\eta\Delta_{-k}$. We use common points $\eta = 1.5$ for proxy-to-semantic edits and $\eta = 1.2$ for semantic-to-proxy edits across seeds, with the same strength applied to learned and shuffled controls of the corresponding direction. Norm-matched random controls are rescaled to the learned direction norm; layer- and module-shuffled controls preserve the update values and scale while disrupting their learned parameter assignment.

Table A4: Data-space rewrite design. Each seed–direction–control–budget combination is repeated over 30 independently sampled subsets.

Item	Specification
Budgets	$N \in \{20, 40, 60, 80, 120\}$ identifying examples within an 8,000-example continuation
Controls	Balanced cue conflict; random-IID cue conflict; random labels; aligned only
Replicates	30 subset draws for each of seeds 42–44 and both edit directions
Ordering	Approximately uniform insertion through the aligned continuation
Optimizer	AdamW, learning rate 2×10^{-4}, same LoRA continuation family as the source model
Switch criterion	Target-signed SPP ≥ 0.8 and ID accuracy ≥ 0.99
Reliable N^*	Smallest balanced-conflict N with switch probability ≥ 0.90 and Wilson lower bound ≥ 0.80 in both directions

A.5 Compute and Provenance

Evaluation uses greedy or direct-logit decoding as appropriate for the short-label task. The anonymized repository centralizes manuscript parameters in `configs/paper.yaml` and `configs/experiments.yaml`, includes a complete command matrix together with result tables and generated figure/table snapshots. `scripts/reproduce_paper.py` rebuilds quantitative artifacts and provenance metadata, and `scripts/validate_release.py` checks expected files, schemas, and hashes.

B Identification: Behavioral Replication and Objective Dependence

B.1 Cross-Setting Endpoint Replication

This section collects endpoint-level replication evidence across changes in task construction, model scale, adaptation method, and architecture. Table A5 summarizes the pure semantic/proxy endpoints, while Fig. A1 tests whether intermediate initial-objective mixtures produce a graded rule continuum across semantic-target shares.

Table A5: Endpoint replication across settings. Means over seeds 42–44.

Setting	ID Acc.	Semantic SPP	Proxy SPP	Mean $\lvert I_{\text{cause}} \rvert$
EvidenceSelect 1.5B LoRA	100.0	$+1.00$	-1.00	0.986
EvidenceSelect 3B LoRA	100.0	$+1.00$	-1.00	0.987
EvidenceSelect 1.5B full FT	100.0	$+1.00$	-1.00	0.982
TokenMatch 1.5B LoRA	100.0	$+1.00$	-1.00	0.976
Phi-3-mini LoRA	100.0	$+1.00$	-1.00	0.978

Table A5 collects the exact endpoint means omitted from the main text. Every setting preserves perfect ID accuracy and opposite rule poles. TokenMatch changes the task generator, Phi-3-mini changes the architecture, Qwen2.5-3B changes scale, and full FT changes the adaptation mechanism. The mixed-objective analysis complements these replications by testing sensitivity to the initial target distribution.

B.2 Mixed Initial-Supervision Continuum

As shown in Fig. A1, completed mixtures lie between the two pure-history poles, indicating a graded dependence on the early objective. Seed coverage is complete for semantic-target shares 0, 0.25, 0.50, 0.75, and 1.00 in the initial phase; the 0.10 and 0.90 conditions contain seed 42 only and are displayed descriptively, with pooled summaries reserved for conditions with complete seed coverage.

Table A6: Seed-level EvidenceSelect endpoints. Every row has 100% ID accuracy and SPP $+1/-1$ for semantic/proxy histories. Entries show I_{cause} on the factorial test and held-out templates as semantic/proxy.

Setting	Seed	Factorial I_{cause} (S/P)	Held-out I_{cause} (S/P)
1.5B LoRA	42	$+0.989/-0.985$	$+0.989/-0.986$
	43	$+0.986/-0.984$	$+0.988/-0.985$
	44	$+0.989/-0.984$	$+0.988/-0.985$
3B LoRA	42	$+0.979/-0.977$	$+0.980/-0.974$
	43	$+0.987/-0.993$	$+0.988/-0.990$
	44	$+0.991/-0.992$	$+0.991/-0.988$
1.5B full FT	42	$+0.985/-0.979$	$+0.985/-0.979$
	43	$+0.985/-0.979$	$+0.985/-0.979$
	44	$+0.987/-0.978$	$+0.987/-0.979$

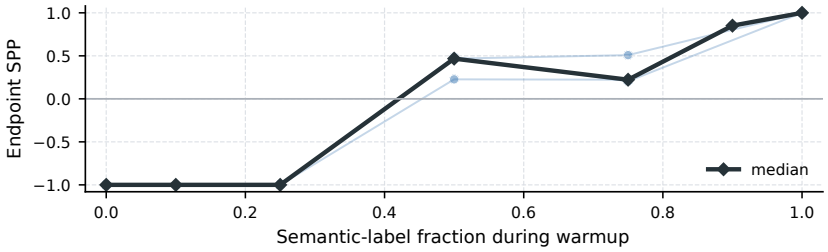


Figure A1: Mixed initial objectives interpolate between endpoint rules. The horizontal axis is the semantic-target share π from Eq. (19). Each line is a completed seed and the dark line is the median; complete three-seed coverage is available for the five central mixture values.

C Diagnosis: Parameter Geometry and Attribution

C.1 ID-Flat Path and Local Objective Sensitivity

The main text reports the behavioral transition along the semantic–proxy interpolation path. Here we give the seed-level path statistics and local objective-sensitivity measurements, then ask how the endpoint difference is distributed across coarse depth regions.

To test whether the endpoint direction is concentrated in one coarse depth region, we partition trainable modules into seven relative-depth blocks and evaluate all $2^7 = 128$ coalitions per seed. A transformer parameter in layer ℓ of an L-layer model is assigned to block

$$b(\ell) = \min\left(6, \left\lfloor \frac{7\ell}{L} \right\rfloor\right). \tag{20}$$

Embedding parameters are assigned to block 0 and the final normalization/output head to block 6. Thus Qwen2.5-1.5B and its full-FT counterpart use four transformer layers per block (0–3, 4–7, …, 24–27); Qwen2.5-3B uses blocks 0–5, 6–10, 11–15, 16–20, 21–25, 26–30, and 31–35. LoRA attribution includes all $q/k/v/o$ target modules in the assigned block; full-FT attribution includes all parameters assigned by the same rule.

For a coalition S, $v(S)$ is the SPP of a hybrid model that takes semantic-endpoint weights in blocks in S and proxy-endpoint weights elsewhere. The exact contribution of block b is

$$\phi_b = \sum_{S \subseteq B \setminus \{b\}} \frac{|S|!(|B| - |S| - 1)!}{|B|!} \left[v(S \cup \{b\}) - v(S)\right]. \tag{21}$$

From the exact SPP Shapley values (Shapley, 1953) we define normalized absolute attribution mass

$$a_b = \frac{|\phi_b|}{\sum_{j \in B} |\phi_j|}, \qquad \sum_{b \in B} a_b = 1. \tag{22}$$

Let $a_{(1)} \geq \cdots \geq a_{(|B|)}$ denote these masses in descending order. The concentration summaries in Figure A2 and Table A8 are the participation ratio (PR), the number of top blocks needed to reach

80% attribution mass (k_{80}), and the largest block share:

$$\mathrm{PR} = \frac{1}{\sum_b a_b^2}, \qquad k_{80} = \min\left\{k : \sum_{r=1}^{k} a_{(r)} \geq 0.8\right\}, \qquad a_{\max} = a_{(1)}. \tag{23}$$

Panel (a) plots a_b; panel (b) plots the cumulative sum $\sum_{r=1}^{k} a_{(r)}$. This provides a coarse depth-level view of where the endpoint difference is supported.

For seed k, let $\Delta_k = \theta_s^{(k)} - \theta_q^{(k)}$, $\hat{\Delta}_k = \Delta_k / \|\Delta_k\|_2$, and let $\lambda_k^\star$ be the first point on the locked interpolation grid whose geometry-validation SPP is nonnegative. The three diagnostic points used by Figure 6c–d and Table A7 are $\lambda_{42}^\star = 0.575$, $\lambda_{43}^\star = 0.550$, and $\lambda_{44}^\star = 0.575$. For an objective $L = (1/N)\sum_i \ell_i$, the unit-direction loss-derivative magnitude is

$$G_L = \left|\nabla_\theta L\big(\theta(\lambda_k^\star)\big) \cdot \hat{\Delta}_k\right|. \tag{24}$$

Figure 6d reports *empirical-Fisher directional curvature*. Operationally, with the fixed step $h = 0.01$, the per-example unit-direction NLL score is

$$u_{i,L} = \frac{\ell_i(\theta(\lambda_k^\star + h)) - \ell_i(\theta(\lambda_k^\star - h))}{2h\,\|\Delta_k\|_2}, \qquad C_L = \frac{1}{N}\sum_{i=1}^{N} u_{i,L}^2. \tag{25}$$

In the differential limit this equals $\hat{\Delta}_k^\top F_{\mathrm{emp}}(L)\hat{\Delta}_k$ with

$$F_{\mathrm{emp}}(L) = \frac{1}{N}\sum_{i=1}^{N} \nabla\ell_i\, \nabla\ell_i^\top. \tag{26}$$

Both G_L and C_L use 256 examples per objective. The norm $\|\Delta_k\|_2$ is the Euclidean norm after vectorizing all merged effective parameter tensors in the endpoint difference (equivalently, the root-sum-square Frobenius norm over those tensors). The aligned loss uses the registered target on an aligned ID split; the semantic and proxy losses use the same balanced complete-16-cell factorial sample with targets s_i and q_i, respectively.

The additional Qwen2.5-7B sensitivity ratio quoted in the main text uses a separate scale-normalized gradient diagnostic. At that model's closest-to-zero-SPP path point, define

$$\widetilde{G}_L = \frac{|\nabla_\theta L \cdot \hat{\Delta}|}{\|\nabla_\theta L\|_{2,\mathrm{supp}(\Delta)} + 10^{-30}}, \qquad \rho_G = \frac{\max(\widetilde{G}_{\mathrm{sem}}, \widetilde{G}_{\mathrm{proxy}})}{\widetilde{G}_{\mathrm{align}} + 10^{-30}}. \tag{27}$$

Here the gradient norm is restricted to parameters on the endpoint-direction support. The reported 13.39–26.64 range is the three-seed range of ρ_G. We reserve G_L for the quantity plotted in Figure 6c; $\widetilde{G}_L$ is used only in this scale-normalized diagnostic.

The main text reports the qualitative path transition and distributed attribution profile. At the seed level, minimum ID accuracies along the interpolation path are 99.8125%, 100.0000%, and 99.9375% for seeds 42–44; the closest-to-zero plotted path points are $\lambda = 0.550$, 0.525, and 0.550. Table A7 gives the seed-level local-sensitivity values, and Table A8 reports the concentration statistics underlying Fig. A2.

Table A7: Directional objective sensitivity at the seed-specific transition point $\lambda_k^\star$. G is Eq. (24) and C is the empirical-Fisher directional curvature in Eq. (25). Factorized ranges span the semantic- and proxy-target losses.

Seed	Aligned G	Factorized G	Aligned C	Factorized C
42	1.13×10^{-6}	46.4–56.4	7.37×10^{-12}	9,154–10,359
43	4.40×10^{-7}	47.1–74.1	1.38×10^{-12}	9,917–13,649
44	1.27×10^{-7}	42.7–51.7	3.41×10^{-13}	8,840–10,024

C.2 Distributed Blockwise Attribution

The summaries in Table A8 agree with Fig. A2: all three families distribute attribution over several blocks. Qwen-3B reaches 80% with three blocks for one seed and four for the other two; all EvidenceSelect and full-FT seeds require four. No single block exceeds 39.4%, which is inconsistent with a one-block explanation under this grouping.

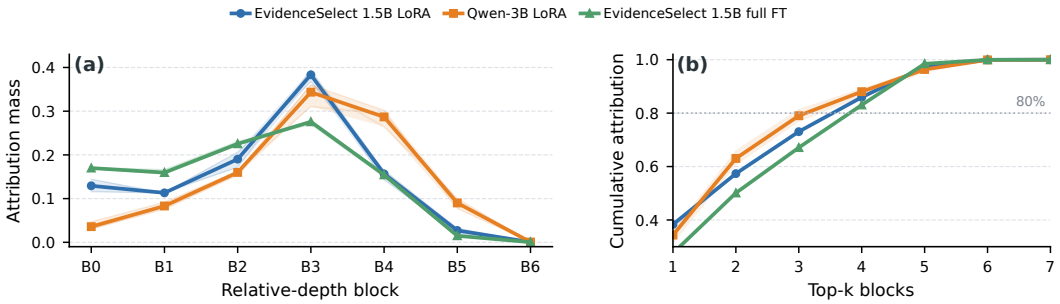


Figure A2: Blockwise causal attribution is reproducibly distributed across depth. Lines are three-seed means and shading is the full seed range. (a) Normalized exact seven-block Shapley mass by relative depth. (b) Sorted cumulative mass: three to four blocks are required to reach 80%.

Table A8: Exact seven-block attribution summaries. Mean [minimum, maximum] across seeds 42–44; k_{80} reports the observed seed range.

Setting	Participation ratio	k_{80}	Largest block share
EvidenceSelect 1.5B LoRA	4.196 [4.137, 4.309]	4	38.3% [36.8, 39.4]
Qwen-3B LoRA	4.121 [3.890, 4.370]	3–4	34.3% [31.1, 36.0]
EvidenceSelect 1.5B full FT	4.874 [4.866, 4.879]	4	27.6% [26.9, 27.9]

C.3 Aggregate Capability Along the Interpolation Path

We evaluate the literal parameter midpoint $\theta_{\text{mid}} = \theta(0.5) = \frac{1}{2}(\theta_q + \theta_s)$ on the AI2 Reasoning Challenge (ARC-Easy/ARC-Challenge) (Clark et al., 2018), HellaSwag (Zellers et al., 2019), Physical Interaction: Question Answering (PIQA) (Bisk et al., 2020), WinoGrande (Sakaguchi et al., 2020), and Massive Multitask Language Understanding (MMLU) (Hendrycks et al., 2021) to check whether traversing the rule-changing path coincides with broad capability degradation. This midpoint is a fixed path location; the local-geometry diagnostic uses the seed-specific $\lambda_k^\star$. Table A9 reports unedited base accuracy and mean midpoint accuracy over completed seeds, with percentage-point changes in parentheses.

Table A9: Base → midpoint accuracy (%). Midpoint values are averaged over completed seeds; parentheses give percentage-point change.

Benchmark	1.5B LoRA	3B LoRA	1.5B full FT
ARC-Easy	75.97 → 75.07 (−0.90)	73.11 → 75.31 (+2.20)	75.97 → 75.37 (−0.60)
ARC-Challenge	46.84 → 46.10 (−0.74)	47.18 → 48.49 (+1.31)	46.84 → 47.07 (+0.23)
HellaSwag	68.25 → 68.37 (+0.12)	73.60 → 74.20 (+0.60)	68.25 → 68.29 (+0.04)
PIQA	75.73 → 76.13 (+0.40)	78.94 → 78.78 (−0.16)	75.73 → 75.91 (+0.18)
WinoGrande	63.22 → 62.80 (−0.42)	68.51 → 69.19 (+0.68)	63.22 → 62.83 (−0.39)
MMLU	60.18 → 59.89 (−0.29)	65.05 → 64.59 (−0.46)	60.18 → 60.05 (−0.13)
Six-task macro	65.03 → 64.73 (−0.31)	67.73 → 68.43 (+0.70)	65.03 → 64.92 (−0.11)

For benchmark b with N_b examples, $\text{Acc}_b = N_b^{-1} \sum_{i=1}^{N_b} \mathbf{1}[\hat{y}_{b,i} = y_{b,i}]$ using the benchmark harness's multiple-choice prediction; the six-task macro is the unweighted arithmetic mean $\text{MacroAcc} = \frac{1}{6}\sum_{b=1}^{6} \text{Acc}_b$. Macro changes range from -0.31 to $+0.70$ points, and the largest displayed decrease is -0.90. The rule transition therefore occurs without a systematic collapse across these aggregate capability checks.

D Control: Rewrite Details and Variability

D.1 Randomized Data-Space Rewrites

The data-space study treats endpoint seed × independently sampled subset as the statistical unit. It contains exactly 3,600 intervention trials: 2 directions × 3 seeds × 5 budgets × 4 controls × 30 subsets. The same schedule and evaluation are used for all controls.

Table A10: Key randomized rewrite outcomes. P→S and S→P denote proxy-to-semantic and semantic-to-proxy rewrites. Each direction/control/budget entry pools 90 trials. The table reports 95% Wilson score lower bounds for the balanced-conflict condition that defines N^*.

Direction	Control	N	Switch	Wilson low	Mean ID
P→S	Balanced conflict	20	68.9%	58.7%	99.991%
P→S	Balanced conflict	40	93.3%	86.2%	99.999%
P→S	Random-IID conflict	40	96.7%	90.7%	99.994%
P→S	Random labels	40	2.2%	–	99.994%
P→S	Aligned only	40	0.0%	–	100.000%
S→P	Balanced conflict	20	95.6%	89.1%	100.000%
S→P	Balanced conflict	40	96.7%	90.7%	100.000%
S→P	Random-IID conflict	40	97.8%	92.3%	100.000%
S→P	Random labels	40	0.0%	–	99.976%
S→P	Aligned only	40	0.0%	–	100.000%

The direction-specific balanced thresholds are 40 examples for P→S and 20 for S→P; the unified reliable budget is therefore $N^* = 40$. At $N \geq 60$, P→S balanced and random-IID conflicts switch every sampled trial, while S→P remains above 94% across the tested range. At the reliable budget, random-IID and balanced conflicts are comparably effective, supporting identifying cue conflict as the operative training signal.

D.2 Fixed-Budget Transfer and Weight-Space Controls

Table A11: Fixed $N^* = 40$ generalization without retuning. Each row pools 30 subset draws; aligned-only controls have zero switches for every family and direction.

Family	P→S switch	S→P switch	Minimum mean ID
Qwen2.5-3B LoRA	96.7%	100.0%	99.994%
Qwen2.5-1.5B full FT	100.0%	100.0%	100.000%

The weight-space controls remain as in the main text: norm-matched random directions (120 draws), layer-shuffled directions, and module-shuffled directions remain at the source rule, whereas all six leave-one-seed-out edits succeed with essentially unchanged ID behavior.

E Runtime Expression: Matched-Input Mediation

This section expands the runtime analysis summarized in Sec. 4.5. We ask where the semantic–proxy endpoint difference is expressed during a forward pass by replacing the complete answer-position state with the paired endpoint's state on the same input and measuring how much of the endpoint-interpolation effect follows that replacement.

Method. For a fixed input, we cache the residual-stream output of transformer layer 21 at the final active prompt token—the answer-selection position—from one endpoint and replace the corresponding state in the paired endpoint before continuing through layers 22–27 and the output head. Layer 21 is selected from a development scan and then evaluated on seeds 42–44, factorial and held-out-template inputs, together with different-label donor, payload-permutation, norm-matched-noise, early-layer, and non-answer-position controls.

For patched, receiver, and donor SPP values, we define the transfer fraction (TF) as

$$\mathrm{TF} = \frac{\mathrm{SPP}_{\mathrm{patched}} - \mathrm{SPP}_{\mathrm{receiver}}}{\mathrm{SPP}_{\mathrm{donor}} - \mathrm{SPP}_{\mathrm{receiver}}}. \tag{28}$$

To connect parameters to computation, we evaluate recipients at $\lambda \in \{0, 0.25, 0.5, 0.75, 1\}$ and impose semantic- or proxy-endpoint states from the same input. Let the total endpoint effect be $\mathrm{TE} = \mathrm{SPP}(1) - \mathrm{SPP}(0)$ on the natural path, and let the direct endpoint effects DE_S and DE_P be the corresponding differences with semantic and proxy states fixed. The mediation fraction (MF) is

$$\mathrm{MF} = 1 - \frac{\max(|\mathrm{DE}_S|, |\mathrm{DE}_P|)}{|\mathrm{TE}| + 10^{-12}}. \tag{29}$$

For a clamp condition c, define the residual parameter-path effect $\mathrm{DE}_c = \mathrm{SPP}_c(1) - \mathrm{SPP}_c(0)$. Clamp blocking is

$$B_c = 1 - \frac{|\mathrm{DE}_c|}{|\mathrm{TE}| + 10^{-12}}. \tag{30}$$

The matched-clamp value is the smaller blocking fraction from semantic- and proxy-endpoint clamps; the orthogonal-control value is the corresponding blocking fraction after a random vector is projected off the matched donor-minus-recipient state difference and rescaled to the same norm. For a set of replacement controls $\mathcal{C}$, the control gap quoted in the text is $\mathrm{Gap} = \mathrm{TF}_{\mathrm{matched}} - \max_{c\in\mathcal{C}} \mathrm{TF}_c$, with the reported value taking the minimum across registered cells. In Figure A3, depth and control panels show the observed min–max transfer range across the registered seed/direction/split cells; the token-specificity panel reports the maximum transfer observed for each tested token position. Panel (c) reports MF from the equation above, matched B_c, and orthogonal B_c.

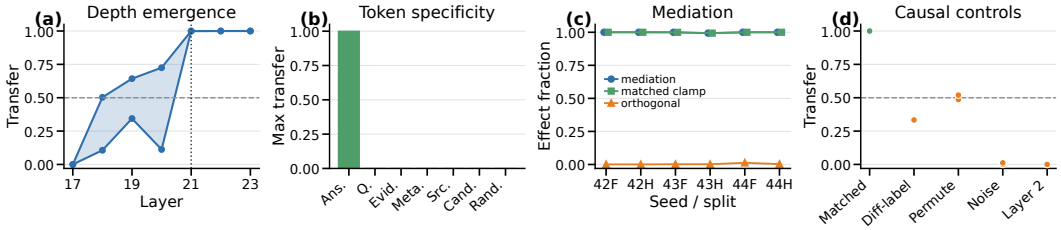


Figure A3: A late matched-input state mediates the parameter coordinate. F/H denote factorial-test/held-out-template splits. (a) Full transfer emerges by approximately layer 21. (b) Tested non-answer positions have zero transfer at layer 21. (c) Mediation and matched-clamp blocking are at least 0.9933, whereas orthogonal clamps block at most 0.0127. (d) Matched replacement exceeds donor, permutation, noise, and early-layer controls.

Result. At the selected post-layer-21 answer-position residual state, matched replacement transfers endpoint behavior on factorial and held-out-template inputs for seeds 42–44: transfer fractions are 0.995–1.000, matched OFR is zero, matched ID accuracy is 1.000, and the minimum matched-minus-best-control gap is 0.480 (Fig. A3a,d). The effect is position specific. Replacing the final question token, semantic-evidence token, proxy-metadata token, selected source label, candidate-final tokens, or a random non-special token produces zero transfer, while answer-position replacement reaches 0.999–1.000 (Fig. A3b).

The same state blocks the parameter-path effect. Across every interpolation value, imposing the semantic endpoint state yields $\mathrm{SPP} = +1$, while imposing the proxy state yields $\mathrm{SPP} \le -0.9867$. Fixing the state removes nearly all residual parameter dependence: mediation fractions and matched-clamp blocking are 0.9933–1.000, compared with at most 0.0127 blocking for norm-matched orthogonal clamps (Fig. A3c). We refer to this as a *late answer-selection state*; the intervention is defined on matched inputs along the studied endpoint path.

Table A12: Activation-state confirmation and mediation. Transfer is the transfer fraction (TF), ID is in-distribution accuracy, mediation is the mediation fraction (MF), and Clamp/Orthogonal report the blocking fraction B_c for matched and orthogonal state clamps. Values are worst cases across seeds and edit directions.

Split	Transfer ↑	OFR ↓	ID Acc. ↑	Mediation ↑	Clamp ↑	Orthogonal ↓
Factorial	0.9996	0.000	1.000	0.9996	0.9996	0.0127
Held-out template	0.9950	0.000	1.000	0.9933	0.9933	0.0029

Replacement controls use different-label donors, globally permuted payloads, norm-matched noise, and an early layer. Their transfer fractions are 0.333, 0.486–0.520, 0–0.012, and 0, respectively, versus 0.995–1.000 for matched donors. Position controls replace the final question token, semantic-evidence token, proxy-metadata token, selected source label, candidate-final tokens individually and jointly, and a random non-special token; every tested non-answer condition has zero transfer at layer 21.

F NATURAL-TEXT AND FULL-CORPUS RETRIEVAL EXPERIMENTS

This appendix reports the natural-text paired histories, the source-disjoint full-corpus retrieval intervention, the common-reader cascade, and the candidate-set disagreement audit used in the main text.

F.1 PAIRED-HISTORY CONSTRUCTION AND ENDPOINTS

The natural-text extension is a source-disjoint short factual source-selection task. One passage contains the supporting fact, while ordinal retrieval metadata independently recommends a source. Semantic and proxy initial phases use the same rendered rows but different targets, followed by the same aligned stream. Qwen2.5-3B is the main model and Phi-3-mini an architecture replication. The Qwen2.5-3B 64k condition is eight passes over the same 8k shared set; shared-example presentations are therefore the training-dose unit.

Table A13: Natural-text paired-history endpoints. Every listed endpoint has ID accuracy 1.000; each row summarizes six endpoints (three seeds × two histories).

Model / shared presentations	Semantic SPP	Proxy SPP	Endpoints
Qwen2.5-3B, 8k	+1.000	−1.000	6
Qwen2.5-3B, 64k	+1.000	−1.000	6
Phi-3-mini, 8k	+1.000	−1.000	6

F.2 FULL-CORPUS RETRIEVAL AND FIXED-READER CASCADE

Corpus and retrieval. The full-corpus pool uses the first 4,000 CounterFact-derived records (Meng et al., 2022) after a deterministic shuffle (seed 8080) that were explicitly excluded from the natural-history data. Each record contributes one support passage and one hard-negative passage as globally addressable documents, giving 8,000 documents and 4,000 queries. Retrieval searches all 8,000 documents with $k = 4$, no query-specific candidate injection, no gold insertion, and no target-model filtering. BM25 (Robertson & Zaragoza, 2009) uses a Lucene-style implementation; dense retrieval uses BGE-large-en-v1.5 (Xiao et al., 2023) classification-token ([CLS]) embeddings with normalized dot product.

Retrieval, routing, and reader metrics. For query i, let g_i be the support document, $r_i(g_i)$ its full-corpus retrieval rank (infinity if absent), $r_i^{(1)}$ the retrieved rank-1 document, and $\hat{g}_i^{(c)}$ the source selected by the model under metadata condition c. Then

$$\text{Recall@}k = \frac{1}{Q}\sum_{i=1}^{Q} \mathbf{1}[r_i(g_i) \le k], \tag{31}$$

$$\text{SCA}^{(c)} = \frac{1}{Q}\sum_{i=1}^{Q} \mathbf{1}[\hat{g}_i^{(c)} = g_i], \tag{32}$$

$$\text{Top1Follow}^{(c)} = \frac{1}{Q}\sum_{i=1}^{Q} \mathbf{1}[\hat{g}_i^{(c)} = r_i^{(1)}]. \tag{33}$$

For the paired true-rank/corrected-rank intervention, the metadata-flip rate reported in Table A15 is

$$\text{Flip} = \frac{1}{Q}\sum_{i=1}^{Q} \mathbf{1}[\hat{g}_i^{(\text{true})} = r_i^{(1)} \ \wedge \ \hat{g}_i^{(\text{corr})} = g_i]. \tag{34}$$

The fixed reader produces a short answer a_i. Normalized exact match is

$$\text{EM} = \frac{1}{Q}\sum_{i=1}^{Q} \mathbf{1}[\nu(a_i) = \nu(y_i)], \tag{35}$$

where $\nu(\cdot)$ lowercases the string, extracts maximal alphanumeric substrings matching `[a-z0-9]+`, and rejoins them with single spaces. Thus punctuation and case are ignored, but no semantic or model-based judge enters EM. The reader qualification's nonempty rate is $Q^{-1}\sum_i \mathbf{1}[\nu(a_i) \neq \emptyset]$.

Table A14: Full-corpus retrieval statistics. Both retrievers search the same 8,000-document corpus with $k = 4$ and no gold insertion. "Aligned means support at rank 1, "Recoverable support at ranks 2–4, and "Miss support absent from the top four. Dense retrieval supplies the recoverable-error set used for the rank intervention.

Retriever	Recall@1	Recall@4	Aligned	Recoverable	Miss
BM25	0.977	0.994	3906	70	24
BGE-large-en-v1.5	0.418	0.992	1672	2296	32

Same-evidence rank intervention. For every recoverable dense query, the passage set and randomized display order are fixed. The *true-rank* condition exposes the actual full-corpus ranks; *corrected-rank* changes only rank metadata so the support is rank 1; *rank-hidden* removes rank metadata; and *swap-back* adds a corrected rank-1 recommendation after the true ranking. Passage-set hashes and non-metadata content hashes are identical across the four conditions. Table A15 reports the 64k endpoint results.

Table A15: Full-corpus routing by seed. All rows use the same 2,296 recoverable dense-retrieval queries and Recall@4$= 1$ within this subset.

History	Seed	True-rank SCA	Corrected SCA	Top-1 follow	Metadata flip
Proxy	42	0.000	1.000	1.000	1.000
Proxy	43	0.000	1.000	1.000	1.000
Proxy	44	0.000	1.000	1.000	1.000
Semantic	42	0.946	0.989	0.044	0.041
Semantic	43	0.986	0.990	0.004	0.003
Semantic	44	0.985	0.987	0.005	0.003

Across all three seeds, true-rank metadata leaves semantic SCA high (0.946–0.986) but drives proxy SCA to 0, while corrected-rank metadata restores both histories to near-perfect support choice. Rank-hidden semantic SCA is 0.981–0.989 and rank-hidden proxy SCA is 0.270–0.278; swap-back SCA is 0.992–0.993 for semantic histories and 1.000 for proxy histories.

Fixed reader. Reader selection uses only question plus oracle-support passage on a disjoint 200-query development set, independently of history-specific selector predictions. Qwen2.5-3B-Instruct reaches 0.960 oracle-support EM with nonempty rate 1.000 and is then used unchanged for every selector history in the cascade.

Table A16: Downstream answer consequence. Values are normalized exact match. Semantic rows show the mean and seed range; proxy outcomes are identical across seeds.

Pipeline	Aligned $N = 500$	Recoverable $N = 2296$
Oracle support	0.944	0.962
Retriever top-1	0.944	0.111
Semantic history	0.935 (0.934–0.936)	0.938 (0.915–0.950)
Proxy history	0.944	0.111

The semantic–proxy recoverable-error EM gap is 0.828 on average and has the same sign in all three seeds, while the aligned-history EM difference is below 0.01. Selected-support rates track the answer gap: semantic histories select the supporting passage on 0.946–0.986 of recoverable queries, while proxy histories never do.

F.3 Controlled Candidate-Set Ranking Conflicts

A complementary retrieval audit uses fixed three-passage candidate sets scored by BM25 or dense retrieval, providing a clean cross-retriever same-evidence replication. For each seed, 500 BM25 and 500 dense top-1 ranking-conflict cases are evaluated under no recommendation, correct recommendation, wrong recommendation, and swap-back conditions. Correct recommendations yield SCA 1.000 for both histories. Semantic histories have wrong-recommendation SCA 0.998–1.000; proxy histories have wrong-recommendation SCA 0 in all six seed–retriever combinations.

For history $h \in \{s, q\}$ and retriever r, define the wrong-recommendation drop

$$\Delta_{h,r} = \mathrm{SCA}_{h,r}^{(\text{no metadata})} - \mathrm{SCA}_{h,r}^{(\text{wrong recommendation})}. \tag{36}$$

Table A17 first averages $\Delta_{h,r}$ over BM25 and dense retrieval for each seed. Its "Excess drop" is $\Delta_q - \Delta_s$, and "Correct-rec. diff." is $\mathrm{SCA}_s^{(\mathrm{correct})} - \mathrm{SCA}_q^{(\mathrm{correct})}$ after the same retriever averaging.

Table A17: Candidate-set ranking-conflict consequence by seed. Values average BM25 and dense ranking-conflict sets. "Excess drop" is the proxy-history minus semantic-history SCA drop under a wrong retrieval recommendation.

Seed	Semantic drop	Proxy drop	Excess drop	Correct-rec. diff.
42	0.001	0.405	0.404	0.000
43	0.000	0.401	0.401	0.000
44	0.000	0.392	0.392	0.000
Mean	**0.0003**	**0.399**	**0.399**	**0.000**

A paired presentation of the same 1,000 queries holds passage text, source IDs, and evidence availability fixed while changing only rank metadata. In the Rank-1 recommendation interface at 64k, semantic wrong-rank SCA ranges from 0.972 to 1.000 and proxy wrong-rank SCA is exactly 0 in every seed–retriever combination; a Top-result phrasing variant and the 8k checkpoints retain the same direction.

F.4 Validation Dilution and Targeted Disagreement Audit

This analysis asks how easily an aggregate validation set can reveal the policy difference measured in Appendix F.3. It is a scenario simulation built from the paired 64k Rank-1-interface outcomes to measure validation dilution under specified disagreement rates. For each seed and each retriever, the source pool contains the same 500 ranking-conflict queries under two metadata conditions. The *aligned condition* assigns the supporting passage the correct rank-1 recommendation; the *disagreement condition* assigns rank 1 to the retriever-selected non-supporting passage. Passage text, source IDs, candidate membership, and query identity are otherwise unchanged.

Paired outcome and detection rule. For query i, let $C_{h,i}^{(m)} \in \{0,1\}$ indicate whether history $h \in \{s,q\}$ (semantic or proxy) is correct under metadata condition $m \in \{c,e\}$ (correct-rank or wrong-rank). We define the paired correctness gaps

$$d_i^{(c)} = C_{s,i}^{(c)} - C_{q,i}^{(c)}, \qquad d_i^{(e)} = C_{s,i}^{(e)} - C_{q,i}^{(e)},$$

so $d \in \{-1,0,1\}$. For every simulated validation sample we compute the paired mean gap $\bar{d}$. A repetition counts as detecting a policy difference when the 95% normal interval

$$\bar{d} \pm 1.96\sqrt{\left(\overline{d^2} - \bar{d}^2\right)/N}$$

excludes zero. Detection probability is the fraction of 10,000 Monte Carlo repetitions satisfying this criterion. This statistic is evaluated independently for all three seeds and both retrievers.

Ordinary IID mixture. We use six scenario disagreement rates and five validation sizes:

$$p \in \{0.001, 0.0025, 0.005, 0.01, 0.02, 0.05\},$$
$$N \in \{500, 1000, 5000, 10000, 50000\}.$$

For each repetition we draw the number of disagreement cases $K \sim \mathrm{Binomial}(N,p)$, then sample K paired outcomes from the empirical $d^{(e)}$ distribution and the remaining $N-K$ from the empirical $d^{(c)}$ distribution. Sampling is performed from the empirical three-point distributions over $\{-1,0,1\}$, so the simulation preserves the observed seed–retriever-specific paired correctness rates. The value of p is deliberately a scenario parameter: these experiments measure dilution as policy-separating cases become rare; production prevalence is outside the scope of this simulation.

Targeted disagreement audit. The targeted design samples directly from the disagreement pool, without replacement, at $N \in \{5,10,20,50,100\}$, and applies the same paired detection rule over 10,000 repetitions. The ordinary-IID comparison for these small budgets uses $p = 0.005$. Thus the comparison changes only the *sampling policy*, while holding the underlying model predictions and detection rule fixed.

Table A18: Detection power for ordinary IID validation and targeted disagreement audits. Values are means over the six seed–retriever combinations. Targeted values at $N = 20$ and $N = 50$ are 1.000 in every individual combination.

Audit design	Scenario p	N	Detection probability
Ordinary IID	0.005	20	0.0000
Ordinary IID	0.005	50	0.0001
Ordinary IID	0.005	100	0.0014
Ordinary IID	0.005	500	0.190
Ordinary IID	0.005	1,000	0.549
Ordinary IID	0.005	5,000	0.908
Ordinary IID	0.005	10,000	0.985
Targeted disagreement	1.0	20	1.000
Targeted disagreement	1.0	50	1.000

These results explain the apparent discrepancy between strong conditional failure and weak aggregate detection. At $p = 0.005$, a 1,000-example IID validation set contains only five disagreement cases in expectation, so its mean detection probability is 0.549 despite a near-maximal conditional semantic–proxy gap on wrong-rank cases. Conditioning the audit on rank–evidence disagreement removes this dilution: 20 targeted cases detect the difference in every tested seed–retriever combination. The conclusion therefore concerns evaluation design—whether the validation distribution contains policy-identifying cases—with production prevalence outside the scope of this analysis.