RESTORING EVIDENCE ROUTING DOES NOT RESTORE EXACT MATCHING IN COMPRESSED KV CACHES

Anonymous authors
Paper under double-blind review

ABSTRACT

Exact lookup requires both access to a record and discrimination of its complete address. We test record recovery, targeted restoration, and source-specific matching under query-agnostic KV-cache compression. Phase 1 reveals a recovery–decision gap in 45 of 144 positive queries under prefix-only attention selection. Both complete source records are independently recoverable from the same cache in these cases, yet the existence readout denies a match. Phase 2 uses independently coded records and equal-byte substitutions of evicted primary-record versus matched nonprimary positions. Primary restoration reduces false absence from 64/96 to 7/96 relative to sham, but also increases near-miss errors. Phase 3 separates the two sources factorially. Restoring A predominantly changes A-queries, and restoring B changes B-queries, yielding a predeclared crossover of 79.17 percentage points with a conservative lower bound of 14.13 points. Joint restoration reduces false absence by 76.04 points while increasing one-field false alarms by 47.92 points. Full and joint restoration both reject all 48 far negatives, yet accept 61 and 55 of 96 near negatives, respectively. This source-selective overmatching separates reinstated source influence from exact Boolean comparison. The resulting diagnostics separately measure record recovery, behavioral fidelity, and query discrimination.

1 INTRODUCTION

Reliable lookup requires an exact decision about the evidence. A compressed language model can answer `NO_MATCH` even when its input database contains a matching record. Such a response asserts absence, whereas `DEFER` expresses uncertainty about the decision. Evaluating these behaviors separately distinguishes erroneous denial from unresolved evidence.

KV-cache policies retain inference state according to recency, attention, block structure, or reconstruction criteria (Xiao et al., 2024; Zhang et al., 2023; Li et al., 2024; Liu et al., 2025; Kim et al., 2025). Yet a retained record can play two different roles: it may make a source recognizable, or it may support checking that *every* queried field matches. Our experiments expose **source-selective overmatching**: restoration makes both an exact address and a one-field near miss more likely to be accepted, chiefly for the restored source. The source-selective response provides a behavioral signature that can be measured independently of record recovery.

Three studies separate recovery, restoration, and source specificity. Phase 1 identifies 45 positive-query cases in which independent readouts recover both complete six-field records from the same compressed cache, while the existence readout denies a match. Phase 2 intervenes on retained positions using new, independently coded records and equal-byte substitutions. Restoring primary-record K/V reduces false absence from 64/96 to 7/96 relative to matched nonprimary substitutions, accompanied by more near-miss errors. Phase 3 crosses the restoration of A and B, testing whether the response follows the restored source (Figure 1).

The decisive result is a $79.17\,\mathrm{pp}$ source crossover, positive in all 24 Phase 3 worlds. A restoration chiefly affects A-queries; B restoration chiefly affects B-queries. The same specificity extends to incorrect acceptance of each source's one-field near misses. Joint restoration reduces false absence by $76.04\,\mathrm{pp}$ while increasing near false alarms by $47.92\,\mathrm{pp}$, both relative to the double-sham control. Restoration reinstates source-specific acceptance of both exact addresses and near misses.

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

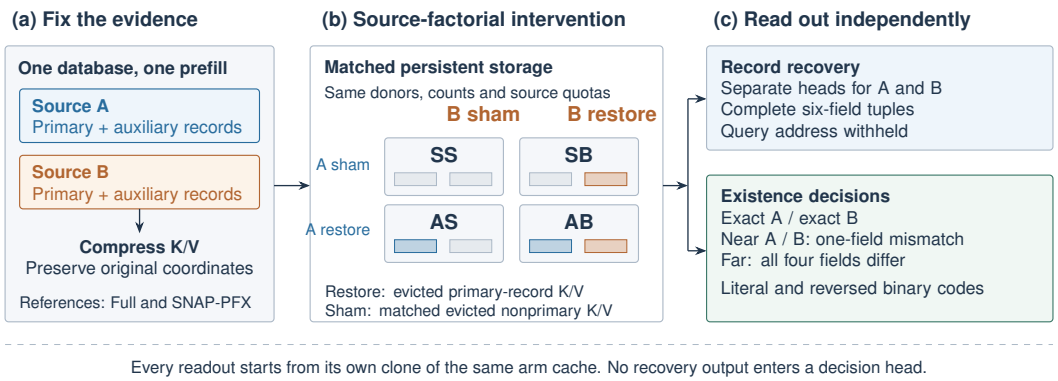

Figure 1: **Restore a source, then test which queries change.** All Phase 3 arms use the same database and original prefill. SS/AS/SB/AB share donor positions; each fixed cache independently serves recovery and existence readouts, with no cross-head answers. Full and SNAP-PFX are separate references. Exact/near query pairs test whether restored source influence is precise. **Sham substitutions preserve persistent bytes, source quotas, and donor removal, isolating primary-state identity from added capacity.**

This tradeoff separates *behavioral fidelity* from *task correctness*. Joint restoration and Full agree on 225/240 literal decisions (93.75%), including 51 shared near-miss false alarms. Both reject all 48 far negatives. The paired decisions show which aspects of the reference's source-conditioned behavior restoration reinstates, including its near-miss errors.

We contribute three complementary diagnostics: a same-cache recovery–decision test, an equal-byte source-factorial intervention, and paired exact/near queries that distinguish source influence from precise matching. Independent cache clones prevent recovery outputs from entering decision readouts. Here *routing* denotes a measured source-selective behavioral response. Phase 1 establishes recoverable false absence; the later cohorts measure the functional consequences of targeted restoration.

2 Related Work

KV retention and available query information. StreamingLLM preserves attention sinks and a recent window (Xiao et al., 2024); H_2O retains heavy hitters (Zhang et al., 2023); Scissorhands exploits persistent token importance (Liu et al., 2023); and PyramidKV varies retention across layers (Cai et al., 2024). SnapKV scores a prompt-tail observation window (Li et al., 2024). FastGen profiles head behavior (Ge et al., 2024), Quest selects pages using the query (Tang et al., 2024), and GemFilter selects input tokens through early layers (Shi et al., 2024). ChunkKV retains chunks, whereas KVzip uses context reconstruction for query-agnostic reuse (Liu et al., 2025; Kim et al., 2025). Compactor combines approximate leverage scores with context-calibrated compression (Chari and Van Durme, 2025); DuoAttention assigns full and limited caches to retrieval and streaming heads, identified through lightweight optimization (Xiao et al., 2025). These are distinct retention and information conditions. Our SNAP-PFX is a prefix-only adaptation scored before any concrete query; CHUNK16 is a 16-token grouping control using those same scores. These controlled adaptations isolate the retention choices examined here.

Long-context retrieval and evidence use. Lost in the Middle studies position-dependent access (Liu et al., 2024); LongBench and RULER evaluate long-context retrieval and reasoning (Bai et al., 2024; Hsieh et al., 2024). SCBench evaluates shared-context workloads across cache generation, compression, retrieval, and loading (Li et al., 2025). Compression studies also examine instruction degradation (Chen et al., 2026), answer–reasoning gaps (Ai et al., 2026), and learned abstention after evidence removal (Khodabandehlou and Krishnamachari, 2026). Our exact/near probes complement workload-level evaluation by distinguishing a factual absence assertion from an inability to decide.

Intervention-based interpretation. Causal tracing, model editing, and circuit analyses study how internal states affect predictions (Pearl, 2009; Meng et al., 2022; 2023; Wang et al., 2022; Geva et al., 2023). Activation-patching conclusions depend on intervention and metric choices (Zhang and Nanda, 2024). We use privileged primary spans to define controlled state substitutions. Their behavioral effect is measured at fixed execution and persistent storage; contextualized K/V may jointly encode content, position, and other features. Equal bytes and common donors hold capacity and donor removal fixed, making primary-state identity the controlled contrast.

3 TASK, CACHE INTERVENTION, AND ESTIMANDS

3.1 A CLOSED-WORLD EXISTENCE TASK

A record is $r = (s, e, p, t, u, v)$: source, entity, property, time, scope, and value. Its address is $\mathrm{addr}(r) = (e, p, t, u)$. For database D and query address q,

$$g(D, q) = \mathbf{1}\{\exists r \in D : \mathrm{addr}(r) = q\}. \tag{1}$$

All four fields must match exactly. Existence depends on the address: two matching records with conflicting values still imply $g = 1$. Executing the predicate on the original database provides the reference labels.

Each Phase 3 world has one primary record per source and auxiliary records in a disjoint entity namespace. Primary A and B differ in every address field. The same latent world is rendered with opaque and readable identifiers; fields are independently coded. Twelve worlds use table syntax and twelve use narrative syntax. Source order and the field changed in near queries are balanced across worlds. Both renderings use artificial identifiers for the same latent records.

For each fixed database, q_A and q_B exactly match its primary records. Queries $q_{\mathrm{near},A}$ and $q_{\mathrm{near},B}$ change one assigned field of the corresponding address, and q_{far} changes all four. Thus each configuration and response encoding has 96 positive, 96 near-negative, and 48 far-negative decisions. The two surfaces and all queries stay within their world for analysis.

3.2 SOURCE-FACTORIAL CACHE SUBSTITUTIONS

We use Qwen2.5-7B-Instruct (Qwen Team, 2024) at a fixed checkpoint with greedy decoding and split-prefix inference. In Phase 3, initial prefixes span 3,072–3,743 tokens. Compressed persistent storage, including original-position indices, occupies 69.5%–70.0% of the corresponding full cache. Original positions and the full-prefix logical clock are preserved. Appendix C specifies retention and memory accounting.

Let K be the retained positions from SNAP-PFX in one layer and KV head. For source $s \in \{A, B\}$, W_s contains eligible evicted positions overlapping its primary-record span, and S_s contains equally many evicted nonprimary positions from the same source. Sham matching prioritizes original-position distance, score distance, and index. The donor set D_s contains the same number of low-scoring eligible retained positions from that source. With $K_0 = K \setminus (D_A \cup D_B)$, the four factorial masks are

$$\begin{aligned} K_{SS} &= K_0 \cup S_A \cup S_B, \quad & K_{AS} &= K_0 \cup W_A \cup S_B, \\ K_{SB} &= K_0 \cup S_A \cup W_B, \quad & K_{AB} &= K_0 \cup W_A \cup W_B. \end{aligned} \tag{2}$$

All arms share donors, retained counts, and source allocations. Inserted K/V comes from the *same original prefill* at its original coordinates. The intervention directly substitutes stored states. Position and attention-score distances define the sham-matching objectives.

Each arm defines a fixed cache state for independent readouts. Two independent heads recover A and B as complete six-field tuples without being given a query address. The literal existence head F receives only its query and required format. CODE01 and CODE10 reverse the physical labels K0/K1 for the same Boolean decision. Phase 3 therefore has $2 + 5 \times 3 = 17$ heads and $24 \times 2 \times 6 \times 17 = 4{,}896$ model readouts in Phase 3. Every head starts from its own clone of that state. Full and unmodified SNAP-PFX are additional references outside the four-arm factorial comparison.

Table 1: **Evaluation targets and readouts.** H and FPR measure explicit erroneous decisions; invalidity and deferral are reported separately. Fidelity to Full is a separate diagnostic, not a goal for false-alarm rate.

Measure	Symbol	Meaning	Direction
False absence	H	Positive query explicitly denied	Lower
Near false alarm	$\mathrm{FPR}_{\mathrm{near}}$	One-field mismatch accepted	Lower
Paired discrimination	Q_{pair}	Exact positive and near negative both correct	Higher
Recoverable denial	J	Both full tuples recovered, positive denied	Diagnostic
Source crossover	T	Own-source versus cross-source response	Positive

3.3 Readouts and evaluation targets

Let $Z_c(q)$ be the parsed decision of arm c. For a valid binary decision define

$$Y_c(q) = \mathbf{1}\{\mathrm{Decoder}_{\mathrm{exist}}(\mathcal{C}_c, q) = \texttt{HAS_MATCH}\} \in \{0, 1\}. \tag{3}$$

$\mathbf{1}\{\cdot\}$ denotes the indicator function. Logical completion bounds retain invalid or unobserved output as unresolved; $Y = 0$ denotes a valid negative decision. `DEFER`, when authorized in Phase 1, is a separate observed action.

With fixed positive set $\mathcal{Q}_+$ and near-negative set $\mathcal{Q}_{\mathrm{near}}$, define

$$H_c = \frac{1}{|\mathcal{Q}_+|} \sum_{q \in \mathcal{Q}_+} \mathbf{1}\{Z_c(q) = \texttt{NO_MATCH}\}, \tag{4}$$

$$\mathrm{FPR}_{\mathrm{near},c} = \frac{1}{|\mathcal{Q}_{\mathrm{near}}|} \sum_{q \in \mathcal{Q}_{\mathrm{near}}} \mathbf{1}\{Z_c(q) = \texttt{HAS_MATCH}\}. \tag{5}$$

For the fixed set of exact/near pairs $\mathcal{P}$, paired discriminability is

$$Q_{\mathrm{pair},c} = \frac{1}{|\mathcal{P}|} \sum_{(q^+, q^-) \in \mathcal{P}} \mathbf{1}\{Z_c(q^+) = \texttt{HAS_MATCH} \wedge Z_c(q^-) = \texttt{NO_MATCH}\}. \tag{6}$$

This form explicitly requires two valid correct decisions. It reduces to $\mathbf{1}\{Y_c(q^+) = 1 \wedge Y_c(q^-) = 0\}$ only when both binary outputs are valid.

Let $R_{A,c}$ and $R_{B,c}$ indicate exact six-field recovery from independent clones of the same cache. The motivating joint rate is

$$J_c = \frac{1}{|\mathcal{Q}_+|} \sum_{q \in \mathcal{Q}_+} \mathbf{1}\{R_{A,c} = 1,\ R_{B,c} = 1,\ Z_c(q) = \texttt{NO_MATCH}\}. \tag{7}$$

Its denominator contains *all* positive inputs, including those with incomplete recovery.

The routing crossover for world w and surface ℓ is

$$T_{w\ell} = \tfrac{1}{2}\left[Y_{AS}(q_A) - Y_{SB}(q_A) + Y_{SB}(q_B) - Y_{AS}(q_B)\right], \quad \widehat{T} = \frac{1}{24} \sum_{w=1}^{24} \frac{T_{w1} + T_{w2}}{2}. \tag{8}$$

A common, query-independent additive AS–SB shift cancels in $T_{w\ell}$. The paired near negatives then assess the precision of that source-selective response.

3.4 Sequential studies and world-level inference

Each phase uses 24 new scientific worlds and a protocol fixed before its outputs. Phase 1 has four material variants and both source orders. Phase 2 introduces fixed-database multi-query tests and independently encoded fields. Phase 3 introduces the crossed restoration and a near-negative control for each source. Each sequential, outcome-informed study addresses its own question and is analyzed separately. Their exact query sets and denominators are listed in Appendix A.

The four predeclared Phase 3 contrasts are T, $Q_{\mathrm{pair},AB} - Q_{\mathrm{pair},SS}$, $H_{SS} - H_{AB}$, and $\mathrm{FPR}_{\mathrm{near},AB} - \mathrm{FPR}_{\mathrm{near},SS}$. Inference uses 24 world-level contributions, preserving dependence among surfaces,

Table 2: **Phase 1: recoverable false absence and heuristic baselines.** All rows use 24 worlds. Entries are counts: A/B recovery and overall correctness have denominator 192; H and J have the fixed 144-positive denominator. J requires both complete six-field records to be recovered from the same cache.

Method	A exact /192	B exact /192	False absence /144	Joint J /144	All correct /192
FULL	136	150	27	16	126
SNAP-PFX	135	134	100	45	75
CHUNK16	146	141	42	24	120
RANDOM	123	116	38	20	124

arms, and readouts. A fixed 4,999-resample stratified bootstrap (Efron, 1979) gives an exploratory simultaneous 95% envelope. We also report a conservative Hoeffding sensitivity interval with a union bound over the four comparisons (Hoeffding, 1963). The bound applies to independent bounded contributions under the specified world generator. Invalid outputs retain fixed-denominator completion bounds. All primary F decisions in Phase 3 are valid. Field, joint-recovery, and encoding analyses below are post hoc summaries of these same outputs; they are reported separately from the primary comparison family.

4 RESULTS

4.1 PHASE 1: RECOVERY AND EXISTENCE DECISIONS DISSOCIATE

Record recovery and existence decisions dissociate on identical cache states (Table 2). SNAP-PFX recovers A and B in 135/192 and 134/192 attempts, yet denies 100 of 144 positive existence queries. Forty-five events satisfy the stronger same-cache joint condition J, compared with 16 for Full. The predeclared difference is 20.14 pp, with exploratory simultaneous envelope $[2.78, 37.50]$ pp. These joint events directly establish the recovery–decision dissociation.

The retention rule changes the frequency of recoverable denial. CHUNK16 has 42 false absences and 120/192 correct decisions, versus 100 and 75/192 for SNAP-PFX; random retention has 38 and 124/192. These Phase 1 comparisons locate the larger false-absence rate in the evaluated prefix selector rather than in cache reduction alone. The subsequent phases use new cohorts and interventions.

The recovery–decision gap also occurs with an explicit uncertainty response. With a literal `DEFER` option, SNAP-PFX and Full have 46 and 17 joint events. Coded outputs yield distinct invalidity and deferral patterns. In the long setting, Full correctly rejects 9/48 genuinely empty queries. Gold lookup labels come from the database predicate; Full supplies the model-behavior reference.

4.2 PHASE 2: PRIMARY RESTORATION CHANGES DECISIONS, WITH A NEAR-MISS COST

The independently coded Phase 2 cohort tests whether an equal-byte change in retained *positions* affects decisions. Before restoration, SNAP-PFX yields 0/48 exact recoveries for each source. The cohort measures how restoring primary states changes decisions from this initial recovery condition.

Relative to matched sham, joint primary restoration reduces false absence from 64/96 to 7/96 and increases the phase-specific QPAIR score from 30/96 to 59/96 (Table 3). A descriptive paired decomposition identifies 57 corrected positives and no reverse positive changes, alongside 27 newly incorrect negatives. All 27 are one-field near misses; far negatives remain correct. The result motivates Phase 3's separate restoration of A and B and its symmetric near-negative pairing. The Phase 2 QPAIR combines exact-A/near and exact-B/far pairs; Phase 3 evaluates the near-only pairing.

4.3 PHASE 3: SOURCE-SELECTIVE OVERMATCHING

Restoration recovers reference behavior, including errors. Joint restoration AB produces 55/96 near false alarms, compared with 9/96 for SS and 61/96 for Full. Both AB and Full reject all 48 far

Table 3: **Phase 2: equal-byte restoration and all heuristic controls.** Counts use 24 worlds with two surfaces. Each method has 96 positives, 96 negatives and 96 query pairs. Here QPAIR pairs exact A with a near negative and exact B with a far negative; Phase 3 instead uses two near-negative pairs.

Method	False absence /96	QPAIR /96	All correct /192
FULL	9	54	150
SNAP-PFX	65	28	123
CHUNK16	29	40	136
RANDOM	55	31	127
Matched sham	64	30	125
Primary restore	7	59	155

Table 4: **Phase 3: source-factorial results, literal F readout.** The study contains 4,896 model readouts across 24 worlds, two surfaces, six configurations and 17 readouts per cache. Entries are event counts clustered within 24 worlds: each arm has 96 positives, 96 near negatives, 48 far negatives and 48 recoveries per source. QPAIR uses both exact/near pairs. Far-FP is zero except for one AS error.

Arm	False absence /96	Near FP /96	Near QPAIR /96	Far FP /48	All correct /240	A exact /48	B exact /48
FULL	4	61	31	0	175	25	19
SNAP-PFX	77	7	12	0	156	3	0
SS	78	9	9	0	153	2	1
AS	44	31	21	1	164	21	1
SB	38	32	27	0	170	3	18
AB	5	55	36	0	180	24	19

negatives. Full defines the reference behavior, including its near-match errors, while the database predicate defines correctness. A post-hoc paired audit finds **225/240 identical F decisions (93.75%)**: 95/96 positives, 82/96 near negatives, and 48/48 far negatives. Fifty-one near-negative agreements are shared false alarms. The paired evaluation separates the reinstated response pattern from predicate correctness (Appendix F).

Table 4 reports the complete F outcomes. The predeclared restoration contrasts use **SS** as their comparator: $H_{SS} - H_{AB} = 76.04\,\text{pp}$, while $\text{FPR}_{\text{near},AB} - \text{FPR}_{\text{near},SS} = 47.92\,\text{pp}$. Near-query paired discrimination improves by $28.12\,\text{pp}$, reaching 36/96. An always-`NO_MATCH` policy would obtain 60% overall accuracy on this query mixture and zero correct exact/near pairs; paired discrimination directly measures the ability to distinguish those matched cases.

The crossed effect is large and source-specific. The routing crossover is $79.17\,\text{pp}$ (Table 5), with positive contributions from every world. The exploratory envelope is $[67.71, 90.62]\,\text{pp}$; the conservative sensitivity bound remains positive at $[14.13, 100.00]\,\text{pp}$. Source A's descriptive factorial main effect on exact A is $70.83\,\text{pp}$, but its effect on exact B is $-1.04\,\text{pp}$. Source B's effects are $84.38\,\text{pp}$ on exact B and $-2.08\,\text{pp}$ on exact A. The crossed response distinguishes source-specific effects from a common affirmative shift.

Recovery responds to the same source manipulation. AS recovers A in 21/48 cases and B in 1/48; SB recovers B in 18/48 and A in 3/48. Unmodified SNAP-PFX has 3/48 and 0/48. Recovery and judgment therefore respond to corresponding source substitutions, with their joint outcomes examined below.

The same selectivity also drives overmatching. Restoring A raises near-A acceptance by $44.79\,\text{pp}$ and near-B acceptance by $2.08\,\text{pp}$. Restoring B raises near-B acceptance by $52.08\,\text{pp}$ and changes near-A acceptance by $-3.12\,\text{pp}$ (Figure 2). Far effects are about one point: AS makes one far error, while Full, SNAP-PFX, SS, SB, and AB make none. The joint exact/near pattern identifies source-selective acceptance with incomplete address discrimination. Acceptance follows the relevant source even when one queried address field differs.

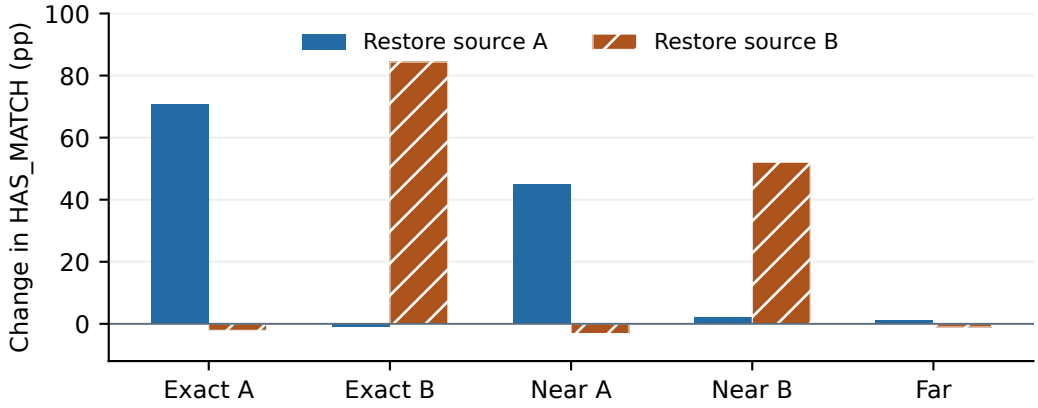

Figure 2: **Restoring A selectively increases both exact-A hits and near-A false alarms; B restoration mirrors this pattern.** Cross-source changes are small, so correct and erroneous acceptance share a source-selective *behavioral signature*. Bars show descriptive factorial effects on `HAS_MATCH`, averaging over the other source factor, both surfaces, and all 24 worlds. Table 5 reports the four predeclared inferential contrasts.

Table 5: **Four predeclared Phase 3 contrasts (percentage points).** Brackets show the exploratory simultaneous 95% bootstrap envelope and the conservative independent-world Hoeffding sensitivity bound. Positive near-FP change is harmful. All estimates use the same 24 worlds.

Contrast	Estimate	Bootstrap envelope	Hoeffding bound
Source crossover T	79.17	$[67.71, 90.62]$	$[14.13, 100.00]$
Q_{pair}: $AB - SS$	28.12	$[16.67, 39.58]$	$[-36.91, 93.16]$
False absence: $SS - AB$	76.04	$[64.58, 87.50]$	$[11.01, 100.00]$
Near FP: $AB - SS$	47.92	$[36.46, 59.38]$	$[-17.12, 100.00]$

Field and readout specificity. The source-selective pattern occurs across all four changed address fields (Table 6). The own-source mean ranges from 25.00 to 70.83 pp, while cross-source changes range from -4.17 to $+4.17$ pp. Each field contributes six worlds; the strata describe variation across fields. Query addresses are checked against the full synthetic database, so each near negative changes exactly one field and matches no record (Appendix H).

Recovery and precise judgment remain distinct within jointly recovered records. Full and AB each recover both records in 5/48 caches. Within these observed subgroups, their ten exact decisions are correct, yet nine and seven near decisions are accepted, respectively. The full recovery–decision table preserves all caches and unresolved formats (Appendix I); the subgroups describe observed recovery states. Under the two reversed code mappings, logical completion ranges remain positive for both exact and near crossovers. For example, the near range is $[46.88, 51.04]$ pp for CODE01 and $[21.88, 42.71]$ pp for CODE10, compared with 48.96 pp for F. The fixed-record ranges enumerate possible completions of unresolved answers (Appendix J).

5 Implications and Scope

Separate access, influence, and precision. Source-selective overmatching shows why evidence access and exact judgment require separate measurements. Equal-byte restoration reinstates source-conditioned hits together with source-conditioned near-miss errors. Paired fidelity identifies the recovered response pattern; matched negatives test whether that pattern respects the complete address predicate.

Interpretive scope. The intervention estimates the behavioral effect of matched-storage substitutions of contextualized K/V. Content, position, and other source-associated features can vary jointly

Table 6: **Source-selective overmatching across address fields.** Literal F, six worlds per field, both surfaces and query sources: 24 near decisions per arm. Own/cross are factorial acceptance changes averaged over the other source factor. Means are post hoc descriptions; Full is the behavioral reference.

Changed field	Own source (pp)	Cross source (pp)	Full near accepts
Entity	+70.83	0.00	18/24
Property	+43.75	-2.08	14/24
Time	+54.17	-4.17	18/24
Scope	+25.00	+4.17	11/24

within those states. Coarse-grained source familiarity is a compatible hypothesis, not a separately measured mechanism. Full is a behavioral reference rather than an exact-lookup oracle, and its agreement with restoration includes errors.

Experimental and statistical scope. Evidence covers one model, a BF16 split-prefix path, synthetic databases, and the measured storage range. The recorded numerical comparisons concern two split-execution paths; equivalence to monolithic prefill was not tested. Other compression methods and model families are separate extensions. The three phases are sequential studies with distinct recovery and query conditions; they are not pooled as independent replications. Phase 1 supplies complete-recovery counterexamples. Comparing Phase 1 with Phase 2 changes both the recovery condition and query design. World-level inference preserves the dependence among readouts. Bootstrap envelopes are exploratory, whereas the bounded-world sensitivity uses its stated independence assumptions. Field strata and recovery-conditioned groups are post hoc descriptions: their differences do not establish field rankings or mediation. Encoding completion ranges concern unresolved outputs, not sampling uncertainty or encoding equivalence. The full-reference and restored-cache recovery subgroups can also contain different worlds.

6 Conclusion

Recovering a record, making a decision follow its source, and matching its address exactly are distinct capabilities. Our three-phase study observes recoverable false absence, establishes an equal-byte primary-position effect, and isolates a large source crossover. The intervention restores source-specific near-miss errors along with true hits. Source-selective overmatching thus establishes an operational distinction between *restored source influence* and *precise address discrimination*.

Reproducibility statement

The accompanying numerical package provides all 4,896 Phase 3 readouts, the 48 synthetic input representations, fixed bootstrap indices, and CPU code for the primary contrasts and additional diagnostics. Appendices C–D specify cache operations and inference; Appendices H–J give the field, joint, and encoding analyses. These scripts recompute statistics from saved outputs. The released commands operate on recorded readouts. The supplement does not distribute the original model-level cache-manipulation runner or full K/V and logit arrays; these are distinct from the numerical reconstruction provided here.

AI Use Statement

AI tools assisted experimental design, implementation, analysis, interpretation, figure preparation, and writing. Codex assisted the local experimental workflow, and ChatGPT assisted evidence review and manuscript preparation. Synthetic databases and gold labels were produced by executable generators. Final diagrams are editable vector drawings checked against the experimental design. The authors take responsibility for the claims, references, and released materials.

ETHICS STATEMENT

The data are synthetic and involve no human participants or personal records. The intervention is intended for controlled diagnosis of evidence use. Deployment reliability requires application-specific evaluation.

REFERENCES

Mengting Ai, Jingrui He, and Yue Guo. Does accuracy equal evidence? reasoning faithfulness under KV cache compression. *arXiv preprint arXiv:2608.01631*, 2026. URL https://arxiv.org/abs/2608.01631.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508*, 2024. URL https://arxiv.org/abs/2308.14508.

Zefan Cai, Yichi Zhang, Bofei Gao, Yuliang Liu, Yucheng Li, Tianyu Liu, Keming Lu, Wayne Xiong, Yue Dong, Junjie Hu, and Wen Xiao. PyramidKV: Dynamic KV cache compression based on pyramidal information funneling. *arXiv preprint arXiv:2406.02069*, 2024. URL https://arxiv.org/abs/2406.02069.

Vivek Chari and Benjamin Van Durme. Compactor: Calibrated query-agnostic KV cache compression with approximate leverage scores, 2025. URL https://arxiv.org/abs/2507.08143v2.

Alex Chen, Renato Geh, Aditya Grover, Guy Van Den Broeck, and Daniel Mingyi Israel. The pitfalls of KV cache compression. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 41530–41553, 2026. URL https://aclanthology.org/2026.acl-long.1926/.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1): 1–26, 1979. doi: 10.1214/aos/1176344552.

Suyu Ge, Yunan Zhang, Liyuan Liu, Minjia Zhang, Jiawei Han, and Jianfeng Gao. Model Tells You What to Discard: Adaptive KV Cache Compression for LLMs. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/639a9a172c044fbb64175b5fad42e9a5-Abstract-Conference.html.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://arxiv.org/abs/2304.14767.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. RULER: What's the real context size of your long-context language models? In *Conference on Language Modeling*, 2024. URL https://arxiv.org/abs/2404.06654.

Mohammadali Khodabandehlou and Bhaskar Krishnamachari. Compression-aware abstention: Teaching LLMs to refuse when KV-compression masks remove answer evidence. *arXiv preprint arXiv:2608.29934*, 2026. URL https://arxiv.org/abs/2608.29934.

Jang-Hyun Kim, Jinuk Kim, Sangwoo Kwon, Jae W. Lee, Sangdoo Yun, and Hyun Oh Song. KVzip: Query-agnostic KV cache compression with context reconstruction. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://arxiv.org/abs/2505.23416.

Yucheng Li, Huiqiang Jiang, Qianhui Wu, Xufang Luo, Surin Ahn, Chengruidong Zhang, Amir H. Abdi, Dongsheng Li, Jianfeng Gao, Yuqing Yang, and Lili Qiu. SCBench: A KV cache-centric analysis of long-context methods. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2412.10319v2.

Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. SnapKV: LLM knows what you are looking for before generation. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/28ab418242603e0f7323e54185d19bde-Abstract-Conference.html.

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. URL https://arxiv.org/abs/2307.03172.

Xiang Liu, Zhenheng Tang, Peijie Dong, Zeyu Li, Yue Liu, Bo Li, Xuming Hu, and Xiaowen Chu. ChunkKV: Semantic-preserving KV cache compression for efficient long-context LLM inference. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://arxiv.org/abs/2502.00299.

Zichang Liu, Aditya Desai, Fangshuo Liao, Weitao Wang, Victor Xie, Zhaozhuo Xu, Anastasios Kyrillidis, and Anshumali Shrivastava. Scissorhands: Exploiting the persistence of importance hypothesis for LLM KV cache compression at test time. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a452a7c6c463e4ae8fbdc614c6e983e6-Abstract.html.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2202.05262.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.07229.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2 edition, 2009.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Zhenmei Shi, Yifei Ming, Xuan-Phi Nguyen, Yingyu Liang, and Shafiq Joty. Discovering the Gems in Early Layers: Accelerating Long-Context LLMs with 1000x Input Token Reduction, 2024. URL https://arxiv.org/abs/2409.17422.

Jiaming Tang, Yilong Zhao, Kan Zhu, Guangxuan Xiao, Baris Kasikci, and Song Han. QUEST: Query-Aware Sparsity for Efficient Long-Context LLM Inference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 47901–47911. PMLR, 2024. URL https://proceedings.mlr.press/v235/tang24l.html.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. *arXiv preprint arXiv:2211.00593*, 2022. URL https://arxiv.org/abs/2211.00593.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2309.17453.

Guangxuan Xiao, Jiaming Tang, Jingwei Zuo, Junxian Guo, Shang Yang, Haotian Tang, Yao Fu, and Song Han. DuoAttention: Efficient long-context LLM inference with retrieval and streaming heads. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.10819.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2309.16042.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, Zhangyang Wang, and Beidi Chen. H_2O: Heavy-hitter oracle for efficient generative inference of large language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2306.14048.

A PHASE STRUCTURE AND DENOMINATORS

The study comprises three sequential cohorts: observational recovery–decision dissociation (Phase 1), joint primary-record restoration (Phase 2), and source-factorial restoration (Phase 3). Each contains 24 latent worlds, separate from two numerical-validation worlds. Within each phase, the model, generator, prompts, budgets, and primary statistics were fixed before evaluating the new worlds. Later phases were designed after inspecting earlier results.

Phase 1 uses four input variants: agreement, conflicting values at a shared address, a distinguishable address/scope, and no matching record. All contain two primary records. The first three imply a positive existence label, including the conflict variant. Two source orders produce 192 input/order cells per configuration and 144 positives. Six heads consist of A recovery, B recovery, forced literal existence, deferral-enabled existence, a three-code existence head, and a matching vector. Full is measured at short and long lengths; three compressed controls at long length. Template and identifier surface are between-world design factors in this phase.

Phase 2 uses one database per world and paired opaque/readable surfaces. Four queries are evaluated under F, D, CODE01 and CODE10: exact A, exact B, one near-A negative, and one far negative. Two recovery heads make 18 readouts per cache. QPAIR pairs exact A with the near negative and exact B with the far negative. Phase 3 instead supplies a near negative for each source, uses F/CODE01/CODE10, and has $2 + 5 \times 3 = 17$ readouts. Phase 2 measures mixed-difficulty QPAIR, whereas Phase 3 uses near-negative pairs for both sources.

Field coding, query families, and layout change between phases. Recovery is consequently evaluated within each cohort; each phase measures its own specified design.

B PHASE 3 GENERATOR AND READOUTS

Primary records independently assign typed codes for entity, property, time, scope, and value. Primary A and B have distinct codes in every address component. Auxiliary records use an entity namespace that does not match any tested query. Primary records sit within their respective source blocks, surrounded by auxiliary records. Source order and the field changed in each near query are balanced over worlds. Input selection depends on the generator and tokenizer before model evaluation.

Both surfaces encode the same latent database. The readable surface replaces opaque field prefixes with field-specific readable prefixes, without adding external semantic truth. Actual initial prefixes in the Phase 3 evaluation cohort range from 3,072 to 3,743 tokens; the admissible range was 3,072–4,096. Paired surfaces can differ in token count, fragmentation, and original positions. The pairing preserves the latent database while allowing these surface properties to vary.

The shared instruction is reproduced below, with line wrapping added:

> This is a closed-world database task. Every primary and auxiliary record is part of the database. A query matches a record exactly when entity, attribute, time, and scope are all equal; value is not part of the query address. Source labels do not change matching. Any matching record is sufficient for HAS_MATCH. NO_MATCH means that no database record matches. DEFER means that you cannot reliably decide; it is not evidence that a matching record does not exist. All identifiers are case-sensitive opaque labels, including identifiers that look like words.

> Do not use outside facts. For a source recovery request, return that source's primary record in the format RECORD(SOURCE_A,entity,attribute,time,scope,value) or RECORD(SOURCE_B,entity,attribute,time,scope,value), with no explanation. Each question is independent. Output only the format requested by the question.

The text defines DEFER for semantic clarity. Under Phase 3's per-head output rules, DEFER is an unresolved-format response in F or a binary code head.

A recovery asks for the complete primary tuple of one source without supplying its address. The A suffix is:

> Return the complete primary record in SOURCE_A. Output only RECORD(SOURCE_A,entity,attribute,time,scope,value).

The B suffix substitutes SOURCE_B. Existence suffixes provide entity, attribute, time and scope, with no value, primary-record text, or earlier model answer. F requires exactly HAS_MATCH or NO_MATCH. CODE01 maps K0 to HAS_MATCH and K1 to NO_MATCH; CODE10 reverses those mappings. The mappings manipulate the output encoding while preserving the query and source intervention.

All readouts share the same token-identical prefix and start from independent clones of the fixed parent cache with empty suffix state. Cache-identity comparisons and reordered-head evaluations verify isolation of the readouts. Statistical inference retains every readout within its parent world.

C Model configuration and cache storage

The checkpoint is `Qwen/Qwen2.5-7B-Instruct` at revision

`a09a35458c702b33eeacc393d103063234e8bc28`.

The recorded environment is Python 3.12.10, PyTorch 2.8.0+cu128, Transformers 4.57.1, NumPy 2.5.3, and SciPy 1.18.1. The model has 28 layers, 28 query heads, 4 KV heads, and head dimension 128. Runs use batch size one, BF16 eager attention, greedy decoding, and an RTX 5090 with 32 GB nominal VRAM under Windows. All three cohorts use the unchanged BF16 model with native split-prefill inference.

For full-prefix length P, dense K/V data occupy

$$B_{\text{full,KV}} = 2 \times 28 \times 4 \times P \times 128 \times 2 = 57{,}344P \text{ bytes.} \tag{9}$$

For a retained count k per layer and KV head, BF16 K/V and int64 original-position indices occupy

$$B_{\text{compressed}} = 57{,}344k + 28 \times 4 \times k \times 8 = 58{,}240k \text{ bytes.} \tag{10}$$

Persistent storage includes each distinct K/V tensor and its required position metadata. Compressed storage is 69.5%–70.0% of the corresponding full cache in Phase 3; unrounded ratios are retained in the numerical records. Four factorial arms have identical counts and shapes for an input. The ratio measures a single service cache. Temporary buffers, transfers, parent copies, and branch clones contribute separately to execution memory.

SNAP-PFX extracts attention over a 64-position terminal prefix window, averages query heads sharing a KV head, and pools scores with width five. Protected positions and the high-priority terminal window remain available. Score ties use original position; retained positions are sorted before attention. Generated suffix K/V grows in BF16, with no subsequent prefix recompression. CHUNK16 groups the same scores into 16-token blocks, while RANDOM uses its fixed selection rule at a matched storage budget.

For each layer/KV-head/source, character-to-token overlaps identify primary-record positions. Protected or already-retained positions do not require restoration. Donors are low-scoring eligible retained positions within the source. Sham positions are evicted nonprimary positions selected lexicographically by distance to the target, score difference, and original index. Cardinalities and source quotas are exact constraints; proximity and score similarity are matching objectives. The retained-position sets satisfy Equation 2.

Recorded matching balance. The original matcher pairs each eligible primary position with an unused same-source sham, prioritizing absolute original-position distance, then score difference, then index. Source quotas, replacement counts, and common donor sets are exact constraints. Scores have no caliper constraint. Table 7 reports the original pair residuals; the unit is a matched token slot within an input, layer, and KV head.

Table 7: Original primary–sham matching residuals. Distances use uncompressed prefix-token indices. Score differences use the stored selector scores. Percentiles are descriptive across matched slots, with experimental and validation inputs separated.

Inputs	Pairs	Median	95th percentile	Maximum
Absolute position distance (tokens)				
Experimental	113,674	33	63	231
Validation	9,312	34	71	149
Absolute selector-score difference				
Experimental	113,674	8.2552433×10^{-6}	4.0411949×10^{-5}	1.2993813×10^{-4}
Validation	9,312	7.8827143×10^{-6}	4.1007996×10^{-5}	1.1530519×10^{-4}

All 5,824 stored layer/head masks, covering 48 experimental and four validation inputs, preserve the common donor set and per-source retained quotas across the four arms. The five compressed configurations have equal recorded persistent storage within each input. The 11,648 source groups include 3,721 experimental and 297 validation zero-dose groups. Donors are shared removal sets; no one-to-one donor–primary pairing is defined.

Independent-readout and execution comparisons. All 408 recorded head-order pairs have identical token sequences and parsed objects. Of these, 401 pairs have valid parses on both sides; the remaining seven retain invalid status. Numerical head-order logit differences were not recorded because the second replay's logits were not saved. The 68 Full-adapter/native-split generation comparisons report identical full-vocabulary logits at 528 paired positions. The 68 teacher-forced comparisons report identical logits at 436 paired positions, corresponding to 136 path evaluations and 872 path positions. Table 8 separates these measurements.

Table 8: Recorded Phase 3 execution comparisons under BF16 eager inference. Zero logit error is the maximum absolute difference in the archived comparisons. The supplement provides the per-position reductions and per-pair records.

Comparison	Pairs	Paired logit positions	Maximum logit error
Head order	408	Unrecorded	Unrecorded
Full adapter/native split: generation	68	528	0
Full adapter/native split: teacher forcing	68	436	0

These are comparisons of the two specified split-execution paths. Monolithic versus split-full execution was not evaluated for the Phase 3 task. The supplement contains the full saved selection-score vectors, retained-position sets, original matching pairs, and readout comparison records. It reconstructs matching residuals and recorded logit-error summaries; elementwise re-evaluation of the vocabulary-logit differences requires the separately archived arrays.

Table 9: Experimental observations and implementation-validation runs. Each phase has 24 scientific worlds. Implementation-validation runs and teacher-forced comparisons are excluded from the experimental denominators.

Phase	Experimental decodes	Validation decodes	Total decodes	Teacher paths
1	5,760	576	6,336	192
2	5,184	936	6,120	144
3	4,896	884	5,780	136
Total	15,840	2,396	18,236	472

D LOGICAL BOUNDS AND STATISTICAL RECONSTRUCTION

All primary F decisions in Phase 3 are valid, so the four primary identification intervals collapse to point estimates before sampling uncertainty. For invalid or missing outputs, the independent reconstruction enumerates feasible binary completions with NO_MATCH reserved for an explicit valid negative. Linear contrasts respect coefficient signs. Conjunctions such as J and QPAIR obey their full logical definition: a valid wrong tuple cannot become correct under completion, whereas an invalid recovery is unresolved. Every row remains in its fixed denominator.

For example, Phase 3 Full and AB have $J = 0$. SNAP-PFX, SS, AS and SB have completion counts $[0, 27]$, $[0, 15]$, $[0, 2]$ and $[0, 6]$ out of 96. These completion ranges encode the remaining uncertainty in unresolved semantic outcomes. A DEFER action, where allowed in Phase 1, is observed and distinct from explicit denial.

The Phase 3 bootstrap uses NumPy PCG64 seed 2026091303 and 4,999 resamples. It independently samples 12 worlds with replacement from each of the table-format and narrative-format strata, retaining all surfaces, queries and arms of each sampled world. The maximum absolute centered deviation over the four contrasts' lower/upper coordinates yields a common empirical 95th-percentile radius of 0.1145833. Expanding the original means by that radius and intersecting with their logical ranges reproduces Table 5. This procedure gives the exploratory nonstudentized envelope.

For the conservative sensitivity bound, each independent world contribution lies in $[-1, 1]$. Hoeffding's inequality and a union bound over four contrasts give

$$\Pr\left(\max_{1\le j\le 4} |\bar{X}_j - \mathbb{E}\bar{X}_j| > \epsilon\right) \le 8\exp(-24\epsilon^2/2). \tag{11}$$

Setting the right side to 0.05 gives $\epsilon = 0.65033157$. The resulting intervals are $[14.13, 100.00]$, $[-36.91, 93.16]$, $[11.01, 100.00]$ and $[-17.12, 100.00]$ percentage points in the table's order. The crossover and false-absence reduction have positive lower endpoints; the remaining two intervals span zero. The independence assumption concerns the specified world-generation process.

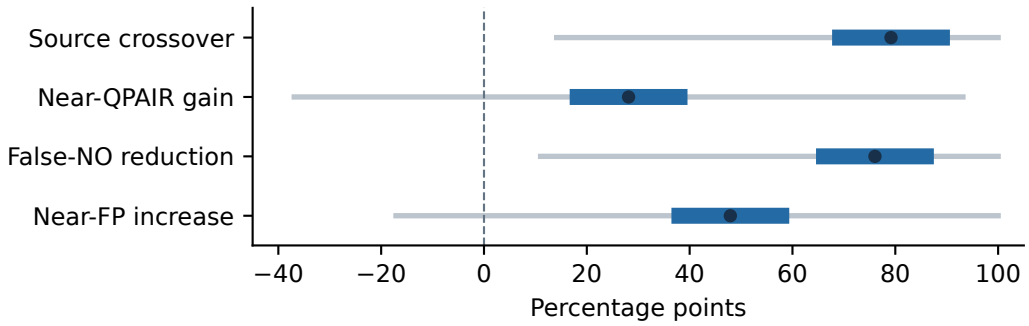

Figure 3: Four primary Phase 3 contrasts. Thick blue: exploratory simultaneous bootstrap envelopes. Thin gray: conservative bounded-world sensitivity intervals, retaining both negative lower endpoints. Points are paired means. Positive near-FP change denotes harm.

E ALL PHASE 3 DECISION ENCODINGS

The complete table retains the actual source/query labels and numeric counts, including all invalid outputs. There are 282 invalid scientific outputs across all Phase 3 readouts, while all primary F outputs are valid. CODE01 and CODE10 are independent suffixes on the same initial cache. The three encodings give separately reported response profiles, with F retained as the primary readout.

F PAIRED FIDELITY TO THE FULL REFERENCE

Behavioral fidelity is a post-hoc comparison of identical parsed F answers, distinct from correctness. For each fixed world, surface, and query, the Full and AB outputs are paired without selecting

Table 10: **Literal and coded readouts retain distinct error profiles.** Complete Phase 3 decisions, rebuilt from all 4,896 scientific outputs. Each arm/encoding has 240 existence decisions; each query column counts HAS_MATCH out of 48. Such answers are correct for Exact A/B and false alarms for Near A/B or Far. Invalidity is separate, never NO_MATCH. Response codes are decoded to semantics before counting.

(a) F Literal HAS_MATCH / NO_MATCH

Arm	Exact A /48	Exact B /48	Near A /48	Near B /48	Far /48	Correct /240	Invalid /240
FULL	46	46	33	28	0	175	0
SNAP-PFX	13	6	6	1	0	156	0
SS	12	6	7	2	0	153	0
AS	47	5	28	3	1	164	0
SB	12	46	5	27	0	170	0
AB	45	46	27	28	0	180	0

(b) CODE01 K0 = HAS_MATCH; K1 = NO_MATCH

Arm	Exact A /48	Exact B /48	Near A /48	Near B /48	Far /48	Correct /240	Invalid /240
FULL	42	40	30	27	1	164	5
SNAP-PFX	12	7	9	2	0	149	3
SS	13	8	8	4	2	150	1
AS	42	7	29	4	0	157	5
SB	14	41	8	29	0	158	7
AB	40	41	28	27	0	164	11

(c) CODE10 K0 = NO_MATCH; K1 = HAS_MATCH

Arm	Exact A /48	Exact B /48	Near A /48	Near B /48	Far /48	Correct /240	Invalid /240
FULL	39	38	19	20	0	171	13
SNAP-PFX	6	3	2	1	0	129	33
SS	7	3	2	1	0	130	33
AS	43	3	20	1	0	151	25
SB	6	42	2	18	0	156	22
AB	39	41	18	14	0	179	14

successful cases. The observed agreement is

$$\widehat{\mathrm{Fid}}_{AB,F} = \frac{1}{240}\sum_{i=1}^{240}\mathbf{1}\{Z_{AB,i} = Z_{F,i}\} = \frac{225}{240} = 0.9375. \tag{12}$$

All relevant F outputs are valid. These 240 paired observations are clustered within 24 worlds; the 240 readouts are repeated measurements.

Table 11 separates agreement on correct decisions from agreement on errors. On exact positives, Full alone is correct in one case. On near negatives, Full alone is correct in four cases and AB alone in ten. In total, Full alone is correct in five cases and AB alone in ten. Of 225 agreements, 170 are correct and 55 are wrong; 51 shared errors are near false alarms. Pairing identifies the exact decisions and errors reinstated by restoration, beyond aggregate error-rate agreement.

Table 11: **Behavioral fidelity includes agreement on errors.** Post-hoc comparison of all Phase 3 F decisions, with no outcome filtering. "Agree" is identical Full/AB answers, not correct answers. The 225 agreements comprise 170 shared correct decisions and 55 shared errors; 51 shared errors are near false alarms.

Query set	N	Agree	Agreement	Both correct	Both wrong
Exact positives	96	95	98.96%	91	4
Near negatives	96	82	85.42%	31	51
Far negatives	48	48	100.00%	48	0
All queries	240	225	93.75%	170	55

G INTERPRETATION OF THE SOURCE CROSSOVER

For one world/surface, suppose the AS–SB difference in the HAS_MATCH indicator or probability is a query-independent additive quantity b. Then the exact-A difference is b and the exact-B difference is also b, so Equation 8 equals $(b - b)/2 = 0$. The large observed positive crossover rejects this specific common-shift description. A source-conditioned familiarity rule, however, could change both the exact and near query of the restored source. It remains compatible with the observed pattern.

The intervention estimates the behavioral effect of replacing the selected coordinates under fixed execution. Because contextualized K/V jointly represents source content, position, and other features, the crossover measures the resulting source-selective influence. Exact and near queries separately assess whether that influence supports the full address predicate.

H FIELD-RESOLVED SOURCE EFFECTS

The post-hoc analysis maps every near query to its differing address component by comparing the original input representation with the corresponding primary record. All 96 near queries differ in exactly one field and match no primary or auxiliary record. Each field has six worlds, with three table-format and three narrative-format worlds. Each source contributes 12 world/surface observations per field; pooling A and B yields 24 decisions per arm. The input field named `attribute` is denoted property in the manuscript.

For a fixed literal query q, let the factorial changes associated with source A and B be

$$e_A(q) = \tfrac{1}{2}\{Y_{AS}(q) - Y_{SS}(q) + Y_{AB}(q) - Y_{SB}(q)\},$$
$$e_B(q) = \tfrac{1}{2}\{Y_{SB}(q) - Y_{SS}(q) + Y_{AB}(q) - Y_{AS}(q)\}.$$

For a query derived from A, own-source and cross-source changes are e_A and e_B; the roles reverse for a B-query. We average within the pre-existing field groups, preserving both surfaces and query sources. Tables 12 and 13 retain source-specific and separate surface/template summaries. These different-world strata provide descriptive means of field heterogeneity.

Table 12: Post-hoc literal-F near effects separated by query source. Each field/source has six worlds, two surfaces and 12 near decisions per arm. Property is the input field named attribute. Values are descriptive means within the six-world field strata.

Field	Query source	Own (pp)	Cross (pp)	Full accepts
Entity	A	+66.67	+0.00	10/12
Entity	B	+75.00	+0.00	8/12
Property	A	+37.50	-4.17	8/12
Property	B	+50.00	+0.00	6/12
Time	A	+41.67	-8.33	9/12
Time	B	+66.67	+0.00	9/12
Scope	A	+33.33	+0.00	6/12
Scope	B	+16.67	+8.33	5/12

Table 13: Separate marginal descriptions of near acceptance under literal F. Both sources are included. The two surface rows use the same 24 worlds; the template rows each use 12 different worlds. These are paired surface or design-stratum descriptions, not prompt or stochastic-decoding robustness tests.

Marginal group	Worlds	Decisions/arm	Own (pp)	Cross (pp)	Full accepts
Opaque	24	48	+45.83	+0.00	27/48
Readable	24	48	+51.04	-1.04	34/48
Table format	12	48	+41.67	-2.08	27/48
Narrative format	12	48	+55.21	+1.04	34/48

I JOINT RECOVERY STATES AND INDEPENDENT DECISIONS

For each world, surface, and arm, both recovery heads and the five literal existence readouts are joined by their recorded common initial cache. Recovery has three states: complete and correct (C), syntactically valid but wrong (W), and unresolved format (U). The joint summaries retain all 48 caches and all 96 exact, 96 near, and 48 far decisions per arm. Unresolved recovery is retained as its own state.

Table 14 reports exact recovery and the original all-positive J completion range. Table 15 retains the full 3×3 recovery-state cross-tabulation, including zero cells. A valid wrong tuple cannot become correct under a logical completion; an unresolved format contributes uncertainty to the upper bound of J.

Table 14: Recovery and independent literal decisions, using all 48 caches per arm. A/B and Both count exact six-field recovery. U counts caches with at least one unresolved recovery. J retains its original all-positive denominator of 96; brackets are logical completion counts, not observed successes or confidence intervals.

Arm	A/48	B/48	Both/48	U/48	J count/96
Full	25	19	5	3	[0,0]
SNAP-PFX	3	0	0	31	[0,27]
SS	2	1	0	27	[0,15]
AS	21	1	1	7	[0,2]
SB	3	18	0	9	[0,6]
AB	24	19	5	2	[0,0]

Among the caches with both records correctly recovered, Full and AB each contribute five caches. Their ten exact decisions contain no denial; their ten near decisions contain nine and seven acceptances, respectively. AS contributes one such cache and accepts its two near negatives. The corresponding subgroup is empty for the other three arms. The table conditions on observed recovery states and preserves the arm-specific subgroup memberships. These proportions describe independent readouts within those memberships.

Table 15: Complete literal-F cross-tabulation. C/W/U denote complete-correct, valid-wrong, and unresolved A/B recovery, respectively. Each decision cell is HAS/NO; all F decisions are valid. Exact and near contain two queries per cache; far contains one. The fixed arm totals are 48 caches and 96/96/48 decisions, including zero cells.

Arm	A/B	Caches	Exact HAS/NO	Near HAS/NO	Far HAS/NO
Full	CC	5	10/0	9/1	0/5
Full	CW	20	39/1	21/19	0/20
Full	CU	0	0/0	0/0	0/0
Full	WC	14	27/1	21/7	0/14
Full	WW	6	10/2	7/5	0/6
Full	WU	3	6/0	3/3	0/3
Full	UC	0	0/0	0/0	0/0
Full	UW	0	0/0	0/0	0/0

Arm	A/B	Caches	Exact HAS/NO	Near HAS/NO	Far HAS/NO
Full	UU	0	0/0	0/0	0/0
SNAP-PFX	CC	0	0/0	0/0	0/0
SNAP-PFX	CW	2	2/2	2/2	0/2
SNAP-PFX	CU	1	1/1	0/2	0/1
SNAP-PFX	WC	0	0/0	0/0	0/0
SNAP-PFX	WW	15	4/26	2/28	0/15
SNAP-PFX	WU	4	4/4	0/8	0/4
SNAP-PFX	UC	0	0/0	0/0	0/0
SNAP-PFX	UW	10	2/18	1/19	0/10
SNAP-PFX	UU	16	6/26	2/30	0/16
SS	CC	0	0/0	0/0	0/0
SS	CW	2	2/2	2/2	0/2
SS	CU	0	0/0	0/0	0/0
SS	WC	1	1/1	0/2	0/1
SS	WW	18	6/30	2/34	0/18
SS	WU	5	2/8	1/9	0/5
SS	UC	0	0/0	0/0	0/0
SS	UW	12	2/22	1/23	0/12
SS	UU	10	5/15	3/17	0/10
AS	CC	1	2/0	2/0	0/1
AS	CW	20	21/19	15/25	0/20
AS	CU	0	0/0	0/0	0/0
AS	WC	0	0/0	0/0	0/0
AS	WW	20	22/18	8/32	0/20
AS	WU	5	5/5	4/6	1/4
AS	UC	0	0/0	0/0	0/0
AS	UW	0	0/0	0/0	0/0
AS	UU	2	2/2	2/2	0/2
SB	CC	0	0/0	0/0	0/0
SB	CW	3	4/2	5/1	0/3
SB	CU	0	0/0	0/0	0/0
SB	WC	14	16/12	10/18	0/14
SB	WW	22	28/16	10/34	0/22
SB	WU	0	0/0	0/0	0/0
SB	UC	4	5/3	4/4	0/4
SB	UW	2	2/2	2/2	0/2
SB	UU	3	3/3	1/5	0/3
AB	CC	5	10/0	7/3	0/5
AB	CW	19	37/1	23/15	0/19
AB	CU	0	0/0	0/0	0/0
AB	WC	14	27/1	16/12	0/14
AB	WW	8	14/2	8/8	0/8
AB	WU	2	3/1	1/3	0/2
AB	UC	0	0/0	0/0	0/0
AB	UW	0	0/0	0/0	0/0
AB	UU	0	0/0	0/0	0/0

J SOURCE SPECIFICITY UNDER ALTERNATIVE RESPONSE ENCODINGS

For encoding e, $T^{(e)}$ applies Equation 8 after decoding its response mapping; $T^{(e)}_{\text{near}}$ substitutes the paired near queries. Both average surfaces within the same 24 worlds. Literal exact T remains primary; the other contrasts are post-hoc diagnostics.

Unresolved outputs admit either binary completion. We combine repeated occurrences before optimizing the linear contrast, retaining coefficient signs. Table 16 reports fixed-record logical completion ranges. Their positive lower endpoints retain source specificity under each encoding; differences in magnitude describe encoding dependence. The original four-comparison inference remains in Table 5.

Table 16: Source crossover under each saved response encoding, on the same 24 worlds and 48 world/surface pairs. Ranges are fixed-sample binary completion bounds in percentage points; unresolved outputs retain both possibilities with coefficient signs respected. Near crossover and coded contrasts are post-hoc descriptions; the original literal exact T remains the primary comparison.

Encoding	Query set	Logical lower (pp)	Logical upper (pp)
F	exact	79.17	79.17
F	near	48.96	48.96
CODE01	exact	64.58	69.79
CODE01	near	46.88	51.04
CODE10	exact	66.67	80.21
CODE10	near	21.88	42.71

K DATA AND CODE AVAILABILITY

The numerical release provides Phase 3 readouts, synthetic inputs, source/field mappings, fixed world-bootstrap indices, original matching records and selector scores, and execution-comparison summaries. CPU scripts reconstruct primary statistics, field and joint summaries, coding bounds, and Full–AB fidelity, with checks of parsing, task labels, and recorded cache identity. This is saved-output statistical reproduction. Original core modules and pinned configuration are provided for inspection. The tested commands reconstruct saved-output statistics; fresh neural execution additionally requires the model and orchestration described in Appendix C. The file manifest and manuscript map identify the recorded observations and the statistics reconstructed by each command.