How Corrupted Chain-of-Thought Exemplars Shift Internal Answer Preferences

Anonymous authors
Paper under double-blind review

Abstract

A correct generated answer does not reveal whether a model ignored an incorrect exemplar answer or recovered from its influence. This raises three questions: (i) does reasoning remove the internal effect of an incorrect exemplar answer, (ii) which components carry that effect, and (iii) does that effect track final-answer accuracy? We study multi-hop comparison questions from 2WikiMultiHopQA, replacing only each retrieved exemplar's answer line with either a plausible but incorrect competing answer from its own source question or an unrelated same-category answer, and adding a reasoning–answer factorial that separates the answer-line effect from exemplar reasoning that contradicts it. We measure the target question's gold-versus-foil logit margin before and after reasoning in Llama-3.1-8B-Instruct, with a same-item cue-probe replication in Qwen2.5-7B-Instruct. Reasoning–answer inconsistency is the largest resolved effect in both models, though weaker in Qwen, and in Llama reasoning preserves it mainly through trace-following. Causal interventions identify five layer-15 writer heads and distributed upper-layer attention as shared endpoints for inconsistency and the smaller answer-plausibility effect; a low-dimensional writer signal reaches the layer-16 MLP, and query direction mediates part of the aligned answer-plausibility effect. Neither the corruption-induced margin drop nor its trace-following component detectably predicts factorial generation failure, and two manipulations differing fourfold in margin effect produce similar accuracy reductions. These results separate internal answer competition from successful output recovery.

1 Introduction

An LLM can answer correctly even when an incorrect answer in a retrieved chain-of-thought (CoT) exemplar has shifted its internal answer preference. Accuracy alone cannot distinguish a model that ignored the incorrect exemplar answer from one that was influenced by it but still produced the correct output. This ambiguity matters in retrieval-augmented generation, where retrieved CoT exemplars act as in-context demonstrations that supply both a reasoning chain for generating an answer and the answer itself (Lewis et al., 2020; Wei et al., 2022; Xie et al., 2024).

Prior work shows that corrupted demonstrations can change in-context behaviour (Min et al., 2022; Yoo et al., 2022; Wei et al., 2023; Halawi et al., 2024). We first introduce a reasoning–answer factorial to establish what replacing an exemplar's answer line actually changes. We then ask three questions that cannot be resolved from output accuracy alone: does query-specific reasoning remove the effect, which components causally carry it, and does it track final-answer accuracy?

To this end, we use multi-hop comparison questions with two candidate answers: the correct answer, which we call the *gold*, and the target question's incorrect competing option, which we call the *foil*. Each prompt contains target evidence and multiple retrieved CoT exemplars. Across exemplar conditions, we remove, redact, or replace only their reasoning chain and final-answer lines, isolating which part carries the effect. The *clean exemplar* condition leaves each exemplar unchanged. *Plausible corruption* replaces its answer with the incorrect option from its own source question, whereas *unrelated corruption* uses an unrelated answer of the same semantic category, such as another person, country, or date. A *permuted control* permutes the plausible-corruption answers among exemplars, so the same answers appear but no longer with their source questions. A *reasoning–answer factorial control* independently varies which of the exemplar's two answers its reasoning derives and

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.

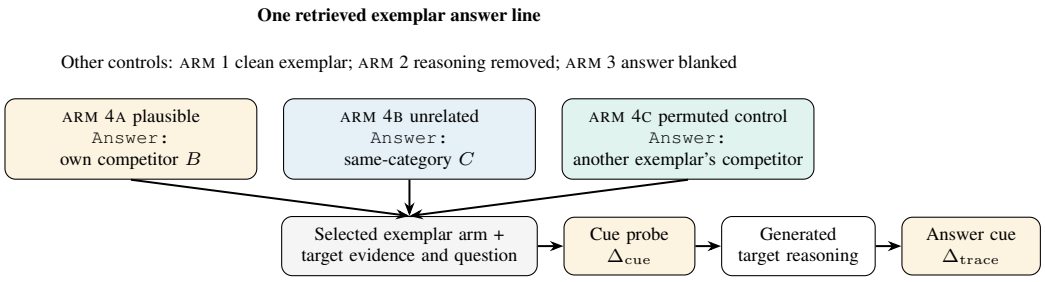

Figure 1: Controlled prompt design. Only exemplar reasoning or final-answer lines change across arms. The cue probe measures the target question's gold-minus-foil logit margin immediately after the prompt. The trace-conditioned probe measures the same margin after teacher-forcing target reasoning (own or common trace) to its answer cue. The permuted control is token-aligned with plausible corruption, which allows interventions that swap a single component of a head's computation (attention pattern versus value, query versus key).

which its answer line states, separating a wrong answer line from *reasoning–answer inconsistency*, when the two disagree. We read gold–foil preference at two points: before the model reasons and after its own reasoning. The first shows whether a corrupted exemplar shifted the preference; the second shows whether reasoning undoes that shift (Figure 1).

Our contributions are:

1. **Reasoning–answer inconsistency dominates the corruption effect.** In Llama, reasoning that derives one answer while the answer line states the other lowers the target margin by 0.946 logits, about four times the answer-line main effect. The effect replicates in Qwen, negative in both models but roughly one-third the size (-0.281 versus -0.946). Relative to each model's clean margin, the gap is wider: -0.033 against -0.228. Qwen's smaller main effects are unresolved rather than demonstrably absent. The corrupted-arm contrasts subtract this shared effect and measure the smaller answer-plausibility component.
2. **Five layer-15 heads write a low-dimensional corruption signal.** Five layer-15 writers carry 72% of the aligned gap (§5) and 66% of inconsistency, and the manipulated answer span supplies 84% of their differential write. One direction shared across the factorial pairs carries 80% of the writers' inconsistency effect while remaining nearly orthogonal to the answer contrast. Downstream, the layer-16 MLP receives 40% of the writers' effect, but the five selected upper-layer heads' queries receive 0.2%; upper-layer attention remains a separate distributed endpoint.
3. **Reasoning preserves inconsistency but dilutes answer identity.** The inconsistency effect stays at -0.23 and -0.22 of the clean margin before and after reasoning. Polarity swaps (§6) show that trace-following carries more of it than evidence does. For the smaller between-corruption contrast, most absolute-gap growth is shared with a placebo, leaving a 21–23% excess.
4. **Internal answer competition dissociates from output accuracy.** Within the corrupted conditions, neither the cue margin nor the trace-following component orders failures above successes (AUC 0.500 and 0.499). The inconsistency and answer-line manipulations each reduce accuracy, by nearly equal amounts (-2.0 and -1.9 points), despite a fourfold difference in margin effect. A proportional cost would be about -7.5 points. An output-only evaluation would record these manipulations as near-equivalent.

Section 2 reviews related work and Section 3 describes the prompt construction and readouts. Section 4 identifies what corruption measures, Section 5 tests its components and propagation, Section 6 tests reasoning, and Section 7 tests generated answers.

2 Related Work

Corrupted in-context exemplars. Randomising demonstration labels can have little effect on in-context classification, although sensitivity varies with the task, prompt, and model (Min et al., 2022; Yoo et al., 2022; Wei et al., 2023). Halawi et al. (2024) show that correct and incorrect few-shot

demonstrations diverge at a critical layer and identify late "false induction heads" that copy false demonstrated information. In classification ICL, Wang et al. (2023c) find that demonstration label words aggregate information in shallow layers and are read from the target position in deeper layers. Our answer-line result has a related source–receiver structure, but it arises in natural-language CoT exemplars, and we separate its endpoints with causal interventions. Unlike classification labels, CoT exemplars display a derivation and an answer. Invalid reasoning can retain some benefits of CoT prompting, while noisy or inconsistent rationales can still impair performance (Wang et al., 2023a; Zhou et al., 2024; Ye & Durrett, 2022). We isolate corruption of the exemplar-answer line and test whether the mediating heads match the false-induction signature: attention to demonstrated answer tokens, preference for answer text over the answer cue, and an output-value circuit that promotes the tokens read (Appendix G.2). Reasoning removal and exemplar-answer blanking remain separate controls.

Reasoning as explanation and computation. CoT text need not faithfully report the process that determined an answer (Turpin et al., 2023; Lanham et al., 2023). We study how appending query-specific reasoning changes sensitivity to corrupted exemplars, rather than asking the model to critique and revise an earlier answer (Madaan et al., 2023; Huang et al., 2024). We therefore measure sensitivity, not faithfulness (Jacovi & Goldberg, 2020).

Context conflict and causal intervention. Irrelevant or conflicting context can distract a model or override parametric knowledge (Shi et al., 2023; Xie et al., 2024). Mechanistic studies localise components that favour contextual or parametric sources and intervene on them (Ortu et al., 2024; Jin et al., 2024; Jiao et al., 2026). We instead compare an incorrect answer bound to its own exemplar question with an unrelated category-matched answer, then measure the resulting contrast before and after query-specific reasoning. Counterfactual patching distinguishes represented information from information the model uses (Vig et al., 2020; Geiger et al., 2021; Heimersheim & Nanda, 2024; Zhang & Nanda, 2024), and node- and path-level methods localise where that use occurs (Goldowsky-Dill et al., 2023; Conmy et al., 2023). We report mediated fractions without treating them as exhaustive because component interventions can induce compensation (McGrath et al., 2023). More broadly, retrieval effects need not follow a simple relevance ordering (Cuconasu et al., 2024).

Logit-based readouts and representation geometry. The logit lens and direct logit attribution project residual states or component outputs through the unembedding to expose answer-aligned evidence (nostalgebraist, 2020; Elhage et al., 2021; Wang et al., 2023b). Linear directions can also encode concepts or reusable functions (Park et al., 2024; Todd et al., 2024), but decodability alone does not imply that a representation is causally used. We therefore pair answer-direction geometry with answer-category-matched nulls and interchange interventions.

3 Experimental Setup

3.1 Benchmark and controlled prompt construction

Task and models. We evaluate Llama-3.1-8B-Instruct (Grattafiori et al., 2024) on all 12,576 2WikiMultiHopQA items (Ho et al., 2020). Mechanistic analyses use 1,251 of 3,040 comparison questions with a correct Llama clean baseline and distinct gold and foil first tokens (Table 3). Qwen2.5-7B-Instruct (Qwen et al., 2025) is evaluated on this Llama-selected set. The five-arm Llama design is also tested on HotpotQA (Yang et al., 2018).

Retrieval. Evidence contexts use KG^{2}RAG (Zhu et al., 2025), which supplies knowledge-graph-connected evidence suited to multi-hop comparison questions; because every arm sees the same context and exemplars, retrieval cannot confound between-arm contrasts. Three exemplars are selected leave-one-out by cosine similarity of L2-normalised `mxbai-embed-large-v1` question embeddings (Lee et al., 2024; Reimers & Gurevych, 2019) from 6,040 exact-match-correct chains. The shared pool and retrieval depth are fixed across models and arms, and neither is tuned. Of the mechanistic targets, 1,233 belong to this pool, making leave-one-out exclusion substantive. The transductive design measures controlled prompt behaviour, not held-out benchmark performance (Appendix A).

Table 1: Exemplar conditions. Exemplar-reasoning removal (ARM 2) and exemplar-answer blanking (ARM 3) are separate controls and are not listed.

Condition	Method	Purpose
Clean exemplar (ARM 1)	Keep each exemplar's reasoning and answer unchanged.	Baseline for the corruption contrasts.
Plausible corruption (ARM 4A)	Replace each exemplar answer with the incorrect competing option from that exemplar's own source question, where one can be parsed.	Measure the effect of a plausible but incorrect answer tied to the exemplar.
Unrelated corruption (ARM 4B)	Replace each exemplar answer with an unrelated incorrect answer from the same semantic category.	Control for answer incorrectness without the exemplar link.
Permuted control (ARM 4C, Llama only)	Reassign the plausible-corruption answers across exemplar slots within the same prompt.	Test dependence on the original exemplar–answer pairing; supplies the token-aligned intervention contrast.
Reasoning–answer factorial control	Independently vary whether the exemplar reasoning derives the gold or foil and whether the answer line states the gold or foil.	Separate the answer-line effect from derivation–answer inconsistency.

Prompt arms. Each prompt contains a target question, context, and the three highest-ranked CoT exemplars. The *clean-exemplar* condition (ARM 1) leaves exemplars unchanged. *Plausible corruption* (ARM 4A) replaces each exemplar answer with its own competing option, and *unrelated corruption* (ARM 4B) with a same-category answer; their contrast is the full-population comparison. Two further controls isolate each component's contribution by removing exemplar reasoning (ARM 2) or blanking exemplar answers (ARM 3). The Llama-only *permuted control* (ARM 4C) supplies the token-matched 688-item contrast used for positional and factor-level interventions. Passages, exemplar identity and order, and source reasoning remain fixed. Appendix A documents truncation and construction audits.

Plausible-corruption replacements negate yes/no answers or substitute the other entity in named-entity comparisons. Some exemplar questions offer no competitor to extract; those answer lines instead take an unrelated replacement built as in ARM 4B, drawn independently, so the two arms still differ there. Source-linked competitors fill 2,292 of 3,750 exemplar answer lines (61.1%), and every prompt has at least one. ARM 4A is therefore a mixture rather than a pure plausible-corruption condition. Appendices A, C, and F report the construction accounting, plausibility, and occurrence audits.

3.2 Readouts and inference

We measure $\Delta = \mathrm{logit}(\mathrm{gold}) - \mathrm{logit}(\mathrm{foil})$, equivalently the projection of the final normalised residual state onto the unembedding-row difference $d = W_U[\mathrm{gold}] - W_U[\mathrm{foil}]$ (Elhage et al., 2021; Wang et al., 2023b). We refer to Δ as the gold–foil margin, and to a difference in Δ between two conditions as a *gap*. Each prompt ends at an *answer cue*, the position where the model must begin its answer. The *cue probe* reads Δ there. The *trace-conditioned probe* reads Δ after appending a reasoning trace for the target question, at the cue where that trace's answer begins. We apply it two ways: with each condition's own generated trace, and with one common trace — the clean-exemplar condition's, appended identically to every condition. Of the 1,251 mechanistic items, 1,243 have complete trace-conditioned readings. On this set, Llama's cue gap, plausible minus unrelated corruption, is -0.35; over the 688 permuted-control items, plausible minus permuted is -0.585. Intervention effects below are signed fractions of these *intact* (no-intervention) gaps' magnitude, so one means the full gap.

Unless stated otherwise, intervals are 95% percentile intervals from 10,000 bootstrap resamples over items, paired across conditions, and each reported fraction resamples its numerator and denominator jointly (Appendix B). Causal language is reserved for interventions. Llama decodes greedily wherever text is generated.

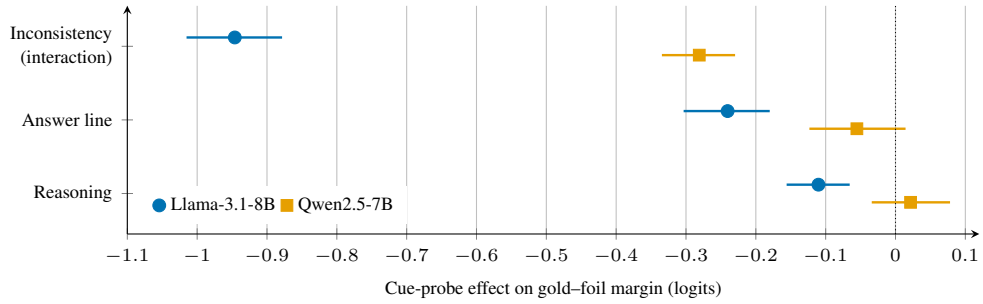

Figure 2: Cue-probe factorial effects on the same 1,221-item manifest, with 95% intervals. Inconsistency is the largest resolved contrast in both models. Qwen reproduces its sign at about 0.30 of Llama's absolute magnitude, although its clean margin is larger. Qwen's answer-line and reasoning main effects are unresolved at this sample size.

4 What Does Answer-Line Corruption Measure?

ARM 4A, ARM 4B, and ARM 4C retain reasoning that derives each exemplar's original answer while changing only the answer lines. They therefore share reasoning–answer inconsistency. To separate this common effect from the stated answer, we cross whether each eligible exemplar's reasoning derives its own question's gold or foil with whether its answer line states gold or foil. Each factor's *main effect* averages over the other; their *interaction* is how much one factor's effect depends on the other's level. The four cells, written reasoning/answer, are the clean-exemplar condition (gold/gold), plausible corruption on eligible slots (gold/foil), a consistent wrong exemplar (foil/foil), and an inconsistent exemplar whose answer remains correct (foil/gold). The construction covers 2,150 of the 2,292 plausible slots across 1,221 items. Appendix D gives the construction checks and results.

The inconsistency term lowers the cue margin by 0.946 logits, about four times the answer-line main effect and eight times the reasoning main effect (Figure 2; Appendix Table 9). It accounts for about 80% of the plausible-versus-clean effect. More sharply, the margin falls by 1.06 logits when every answer line is correct but the reasoning derives the other option. A linear margin-scale analysis attributes about 68% of the inconsistency effect at the mean clean margin to proportional shrinkage and 32% to a constant foil-directed shift. This scale-dependent description does not establish that the model detects inconsistency or infers a rejected-option rule. Because ARM 4A, ARM 4B, and ARM 4C all retain derivations that disagree with their replacement answers, comparisons among them subtract a shared inconsistency and isolate the smaller answer-plausibility effect studied below.

On the same 1,221-item manifest, Qwen reproduces the inconsistency interaction with the same sign and about 0.30 of Llama's absolute magnitude (-0.281 $[-0.335, -0.230]$). Qwen's answer-line and reasoning main effects are unresolved at this sample size. Qwen's clean margin on these items is larger than Llama's ($+8.46$ versus $+4.16$). Its clean gold preference is also slightly higher (88.0% versus 86.4%). The attenuation is therefore not explained by a smaller logit scale, and the Llama-based selection does not hand Qwen a harder item set. Relative to its own clean margin, Qwen's effect is -0.033 $[-0.039, -0.027]$, compared with -0.228 $[-0.243, -0.213]$ for Llama, so normalisation widens rather than closes the cross-model difference. We claim qualitative replication—the same sign and dominant component—with a substantial quantitative difference in sensitivity. Scaling Llama's main effects by 0.30, the observed interaction ratio, predicts Qwen effects of -0.07 and -0.03, both inside Qwen's intervals; we therefore read the unresolved main effects as underpowered rather than absent (Figure 2; Appendix Table 11).

5 Which Components Causally Mediate These Effects?

We first localise the token-aligned plausible-minus-permuted answer-plausibility effect at the cue position, then test whether its layer-15 writers and upper-layer readers transfer to the dominant inconsistency effect. We use *writers* for the five layer-15 heads that introduce it at the cue position, *readers* for all attention in layers 24–31, and *carriers* for five heads selected from them. The writers were selected on the aligned contrast and the carriers on plausible-minus-unrelated; both transfer

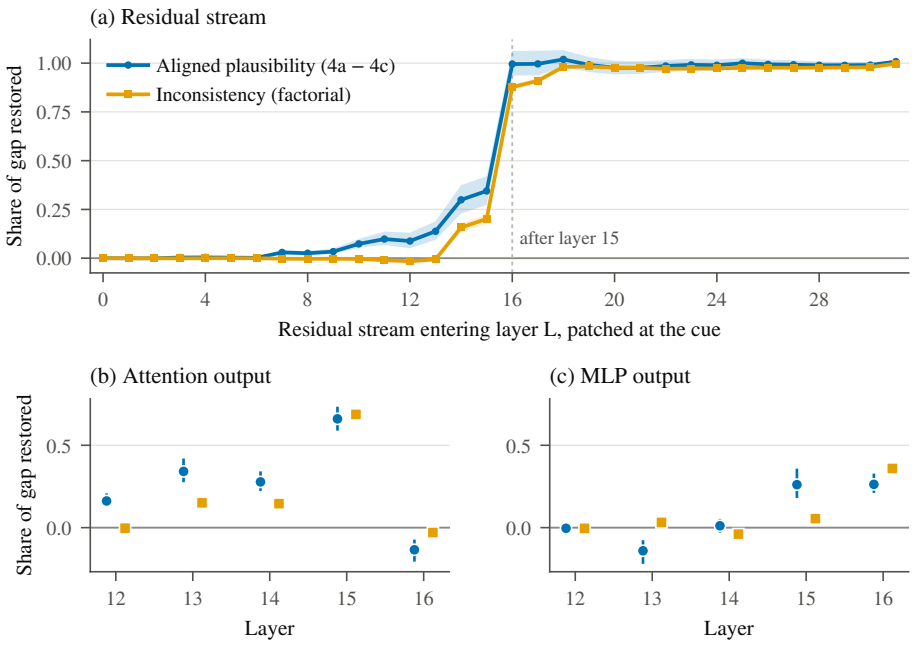

Figure 3: Cue-position arrival of the aligned ARM 4A−ARM 4C difference ($n = 326$) and the factorial inconsistency effect ($n = 584$), both held-out, with 95% joint-bootstrap intervals; most inconsistency intervals are narrower than their markers. (a) Per-item residual interchange; each coordinate is the residual stream entering that layer, so the jump at 16 is written by layer 15. (b, c) Attention and MLP outputs at layers 12–16, where shares may be negative; the layer-16 MLP adds more to inconsistency (0.36) than to the aligned difference (0.26). Effects are bidirectional interventions, not logit-lens readouts.

to inconsistency without reselection. Direction transplant, query factorisation, and the span-level attention signature are tested only on the aligned contrast.

5.1 Writing the aligned difference into the cue position

Layer 15 dominates the arrival step. Interchanging the cue-position residual stream between a contrast's two conditions recovers little of the aligned or inconsistency effect through layer 13, then most of it at the stream entering layer 16 (Figure 3). Layer-15 attention carries 0.66 of the aligned gap and 0.69 of the inconsistency gap. Sublayer shares measure downstream causal influence, not an additive decomposition (Appendix D.5).

Five layer-15 heads carry both writes. On held-out items, H17, H8, H11, H2, and H1 move 0.72 of the aligned gap—1.10 $[0.98, 1.23]$ of layer-15 attention, indistinguishable from all of it—and 0.66 of inconsistency without reselection (Figure 4a). Decomposing their write by source position places 84.1% $[83.2, 85.0]$ of the between-condition difference on the manipulated answer span, despite only 12–16% attention mass there (Appendix D.7).

Their differential write is effectively one-dimensional. One selection-split direction retains 79.8% $[76.2, 83.6]$ of the writers' inconsistency effect on held-out items, whereas a norm-matched random direction retains none. This direction is nearly orthogonal to the gold–foil unembedding contrast, so it does not write directly onto that contrast (Appendix D.8).

The direction feeds the layer-16 MLP, not carrier queries. The layer-16 MLP receives 0.396 of the writers' inconsistency effect, but carrier queries receive 0.002 (keys and values unpatched). The writers' effect therefore does not reach carriers through their queries. Transmission after the MLP remains unresolved (Appendix D.9). Appendix D.11 reports cross-condition specificity.

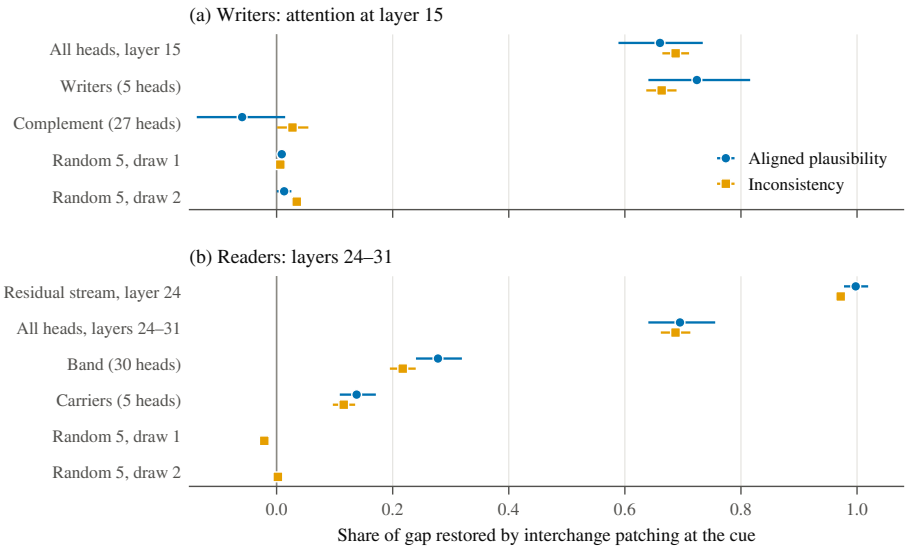

Figure 4: Writer and reader interventions on the aligned contrast and held-out inconsistency items, shown as shares of each intact gap with 95% joint-bootstrap intervals. (a) The five writers carry 0.72 of the aligned gap and 0.66 of inconsistency, matching the whole layer; their complement and random sets carry little. (b) Patching the layer-24 residual stream recovers the full gap: all upper-layer attention carries 0.70 and 0.69, the thirty attribution-selected heads 0.28 and 0.22, and the five carriers 0.14 and 0.12. Random five-head sets were run for inconsistency only. The panels show two causal endpoints rather than a direct route between them; Appendix Table 23 tests that route.

Qwen's residual-stream effect arrives later in relative depth. On 250 items, every Qwen residual-patch interval through layer 17 covers zero. The recovered share rises to 0.318 entering layer 18, 0.700 entering layer 21, and 1.002 entering layer 27. Layer 17 of 28 (0.61 of depth) writes the largest arrival step, against layer 15 of 32 (0.47) in Llama; we compare steps rather than first resolution, because first resolution depends on power (Llama's larger gap resolves shares below 0.01 from layer 3). This comparison is not power-matched: Qwen's intact gap is roughly one quarter of Llama's at the same sample size, and layer-17 attention already resolves at 0.166 $[0.046, 0.265]$ before the residual-stream effect does. We therefore interpret the result as later residual arrival, not later mechanistic onset. At layer 18, attention and the MLP each carry a resolved share, so the Llama writer-head structure cannot yet be claimed for Qwen (Appendix D.3, Table 12).

5.2 Reading both differences in upper-layer attention

Attribution localises a candidate band. Direct logit attribution (Elhage et al., 2021; Wang et al., 2023b) identifies an upper-layer band (layers 24, 25, 27 and 30) and a 17-head core, stable across three item partitions, in Llama. The carriers are the core's five largest-magnitude heads, selected by direct contribution on 651 mechanistic items and patched at the cue position on the 600 held-out items (Appendix G). Attribution supplies candidates, not evidence of causal use.

Upper-layer attention carries most of both cue effects. All attention in layers 24–31 moves 0.70 of the aligned gap and 0.69 of the inconsistency gap. The thirty attribution-selected heads move 0.28 and 0.22. The five carriers move 0.14 and 0.12 (Figure 4). Random five-head sets, run for inconsistency, move negligible shares (-0.021 and $+0.002$). Upper-layer attention is therefore a shared, distributed reader, with the carriers a small resolved subset. Cross-effect comparisons remain descriptive because the populations differ (688 aligned versus 584 held-out factorial items).

Their false-induction resemblance is not copying. Four of the five match the false-induction signature of Halawi et al. (2024), copying an in-context answer, but copy score does not predict per-item gaps and no error generated in the factorial exactly copies an exemplar's answer (Appendices G.2 and J.2).

On the aligned contrast, carriers act through query direction. Patching the head's output, its attention pattern, or only its queries each moves 0.138–0.143 of the gap, whereas key and value patches are indistinguishable from matched controls. Query direction alone reproduces the effect; magnitude alone does not. The item-specific query direction therefore reweights nearly unchanged attended content (Appendix G.3). This factorisation has not been repeated for inconsistency.

6 How Does Query-Specific Reasoning Change the Effect?

We first test whether reasoning removes the dominant inconsistency effect, then analyse the smaller answer-plausibility contrast across controls.

6.1 The inconsistency effect

Reasoning does not remove the inconsistency effect. Under the shared trace, inconsistency grows from -0.946 to -3.047 logits, but stays at -0.23 and -0.22 of the clean margin before and after reasoning. By contrast, the answer-line main effect falls to 0.01–0.02 on the same scale. Each condition's own trace gives the same inconsistency effect, so persistence is not an artifact (Appendices D.13 and the preceding subsection).

Trace-following dominates the persistent effect. We supply the shared trace as written and with its polarity reversed, deriving the other answer. The half-difference of the two margins is *trace-following*; their half-sum is polarity-invariant *evidence*. A consistent wrong exemplar leaves trace-following unchanged, whereas an answer contradicting its reasoning reduces it. Both channels carry inconsistency relative to their clean scales, but trace-following dominates (-0.270 versus -0.127; Appendix D.14).

6.2 The smaller answer-plausibility effect

The remaining analyses compare corrupted conditions, which share this inconsistency.

Plausible corruption lowers the cue margin more than unrelated corruption. The mean Llama gap is -0.35 (Table 34). Its max-t band first lies below zero at layer 16 and again throughout layers 23–32 (Appendix Figure 9a). The band shows where the difference becomes linearly decodable from the residual stream. Decodability is not evidence that the model uses the difference.

Reasoning increases raw contrast magnitude. In Llama, the mean absolute gap rises from 1.070 to 2.101, but the placebo contrast—ARM 4B against a second, independently drawn unrelated corruption—already reaches 0.78 of that cue magnitude and also grows under the shared trace. Individual condition margins expand faster than any contrast, indicating redistribution on an expanding logit scale rather than scale-invariant amplification.

Reasoning redistributes rather than removes the contrast. In Llama, the item-level correlation between cue and trace gaps is modest (0.396) and the signed mean moves from -0.35 to -0.07. Because the mean absolute gap grows, this movement reflects cancellation rather than uniform weakening (Appendix E).

Relevant reasoning produces a modest excess over the placebo. On 684 of the 688 permuted-control items, the plausible-minus-unrelated and plausible-minus-permuted contrasts both grow 21–23% more than the placebo. The excess survives the permuted control. Matched-quantile estimates are non-monotone and imprecise (Appendix Figure 7 and Table 30).

A common trace separates only matched-irrelevant insertion. The cue-to-own change decomposes exactly into three steps—trace insertion, query-relevant content, and arm-specific divergence—of which only insertion resolves. The content interval includes zero and content plus divergence remains descriptive (Table 33). After reasoning, the signed-mean simultaneous band narrows to layers 26–31, which does not establish item-level shrinkage (Appendix E.6).

6.3 PERMUTED CONTROL

The Llama-only permuted control still shows a negative cue gap of -0.585 $[-0.677, -0.495]$. Its absolute gap grows under the common trace, but its signed mean does not approach zero. Signed-mean cancellation is therefore specific to contrasts involving the unrelated arm, while magnitude growth and the excess survive the control change (Appendices A.6–E); no Qwen capture is reported.

7 DO THE INTERNAL READOUTS PREDICT GENERATED ANSWERS?

Corruption-effect readouts do not detectably predict factorial failure. Within the corrupted factorial conditions, neither the cue margin nor the dominant trace-following component separates incorrect generated answers from correct ones (AUC 0.500 $[0.428, 0.572]$ and 0.499 $[0.419, 0.577]$). Aggregate accuracy shows a related mismatch: inconsistency and the answer-line effect are both resolved, yet their accuracy reductions are similar in size at -2.0 $[-2.8, -1.3]$ and -1.9 $[-2.7, -1.1]$ points. Because inconsistency is fourfold larger on the margin, scaling the answer-line reduction predicts about -7.5 points, outside the observed interval. Appendix Tables 70, 71 and 73 report exact-match accuracy, with the same ordering on HotpotQA (Appendix J.3).

As an exploratory observation, the evidence component has the higher failure AUC for every error type, even though trace-following accounts for more of the margin effect. No component's AUC lower bound exceeds 0.5, and all paired differences from the cue margin include zero; this ordering is therefore a follow-up hypothesis rather than a result (Appendix Tables 76 and 77).

Recovery is sparse and not head-specific. Many traces that change under corruption still end in the correct answer. Successes have a larger trace-conditioned margin than failures among complete cases, but both are measured after the corruption, so this comparison conditions on an outcome (Montgomery et al., 2018). The pre-specified comparison of the five carriers against one random control set is null, and ten further draws place that control in the tail of the null distribution. Recovery remains sparse and configuration-dependent (Appendix Tables 82 and 83). The experiment does not establish head-specific control of output recovery.

8 LIMITATIONS

Qwen replicates the inconsistency interaction and yields a later arrival step in relative depth, but the timing comparison is not power-matched and does not establish later mechanistic onset. Qwen's writer heads have not been ranked, and trace-conditioned and generation tests remain Llama-only. None of the cue margin, its dominant trace-following component, or the writers' answer-span write detectably predicts factorial failure. The primary cue-margin test had 80% power to detect AUC ≥ 0.60, so weaker associations remain possible. Establishing a per-item behavioural link requires a persistent writer intervention during generation. The writer direction rules out a direct linear answer readout, not nonlinear decodability; the path test leaves keys and values unpatched, and the effect is absent at the first generated reasoning token. Two further limitations bear on the generation null: the trace-conditioned probe measures a supplied rather than a free-running trace, and the generation population is near ceiling, with 394 of 467 generation items already correct. Both models are 7–8B instruction-tuned; scale is not varied. Appendices J.3 and L discuss token alignment, recovery sparsity, item selection, transductive retrieval, and HotpotQA's behavioural-only replication.

9 CONCLUSION

Answer-line corruption acts mainly through inconsistency between an exemplar's reasoning and stated answer, which Qwen replicates with weaker sensitivity. In Llama, reasoning preserves this effect, largely through trace-following. Five layer-15 writers and distributed upper-layer attention carry both inconsistency and the smaller answer-plausibility effect. The writers' differential write is effectively one-dimensional, and that direction is nearly orthogonal to the gold–foil direction and feeds the layer-16 MLP rather than the carriers' queries. The margin diagnoses gold–foil competition that output accuracy alone does not reveal, but it is not a proxy for accuracy. No corruption-effect readout we test detectably tracks factorial failure.

Reproducibility Statement

Appendix A documents data construction and audits, Appendix B defines the readouts and inferential procedures, and Appendix L records the analysis environment and verification checks. Greedy captures are deterministic within a fixed software environment. Unless otherwise stated, bootstrap analyses use project seed 20260902, with each analysis's seed recorded in the released artifacts; factor-intervention share bootstraps use seed 0, and generation-time control draws use seeds 1000–1009. Greedy inference uses no sampling seed. Recorded-environment checks reproduce Llama exactly and Qwen within fp16 tolerance.

Ethics Statement

This work analyses publicly available models on public question-answering benchmarks and involves no human subjects or personally identifying data. It characterises a failure mode of retrieval-augmented prompting. The findings are diagnostic and do not introduce a method for degrading model behaviour beyond the cited prior work.

AI Use Statement

We used generative AI tools to identify relevant literature and provide feedback on the framing of the paper. The authors reviewed and verified all AI-assisted citations and take responsibility for the final content.

References

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. The power of noise: Redefining retrieval for rag systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, 2024. doi: 10.1145/3626772.3657834.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021.

Edgar C. Fieller. Some problems in interval estimation. *Journal of the Royal Statistical Society: Series B (Methodological)*, 16(2):175–185, 1954. doi: 10.1111/j.2517-6161.1954.tb00159.x.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. Localizing model behavior with path patching. *arXiv preprint arXiv:2304.05969*, 2023.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Danny Halawi, Jean-Stanislas Denain, and Jacob Steinhardt. Overthinking the truth: Understanding how language models process false demonstrations. In *International Conference on Learning Representations (ICLR)*, 2024. URL `https://openreview.net/forum?id=Tigr1kMDZy`.

Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*, 2024.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, 2020.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2310.01798.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4198–4205, 2020.

Difan Jiao, Di Wang, and Lijie Hu. Understanding the dynamics of demonstration conflict in in-context learning. *arXiv preprint arXiv:2603.04464*, 2026.

Zhuoran Jin, Pengfei Cao, Hongbang Yuan, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. Cutting off the head ends the conflict: A mechanism for interpreting and mitigating knowledge conflicts in language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 1193–1215. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.70.

Daniël Lakens. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4):355–362, 2017. doi: 10.1177/1948550617697177.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

Sean Lee, Aamir Shakir, Darius Koenig, and Julius Lipp. Open source strikes bread: New fluffy embedding model. Mixedbread model release, 2024. `https://www.mixedbread.com/blog/mxbai-embed-large-v1`.

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

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023.

Thomas McGrath, Matthew Rahtz, Janos Kramar, Vladimir Mikulik, and Shane Legg. The hydra effect: Emergent self-repair in language model computations. *arXiv preprint arXiv:2307.15771*, 2023.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022.

Jacob M. Montgomery, Brendan Nyhan, and Michelle Torres. How conditioning on posttreatment variables can ruin your experiment and what to do about it. *American Journal of Political Science*, 62(3):760–775, 2018. doi: 10.1111/ajps.12357.

nostalgebraist. Interpreting gpt: The logit lens. LessWrong, 2020. `https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens`.

Francesco Ortu, Zhijing Jin, Diego Doimo, Mrinmaya Sachan, Alberto Cazzaniga, and Bernhard Schölkopf. Competition of mechanisms: Tracing how language models handle facts and counterfactuals. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *International Conference on Machine Learning (ICML)*, 2024.

Belinda Phipson and Gordon K. Smyth. Permutation p-values should never be zero: Calculating exact p-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, 9(1):Article 39, 2010. doi: 10.2202/1544-6115.1585.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 3982–3992. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1410.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, 2023.

Eric Todd, Millicent L Li, Arnab Sen Sharma, Aaron Mueller, Byron C Wallace, and David Bau. Function vectors in large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. Towards understanding chain-of-thought prompting: An empirical study of what matters. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2717–2739. Association for Computational Linguistics, 2023a. doi: 10.18653/v1/2023.acl-long.153.

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. In *International Conference on Learning Representations (ICLR)*, 2023b.

Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. Label words are anchors: An information flow perspective for understanding in-context learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9840–9855, Singapore, 2023c. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.609. URL https://aclanthology.org/2023.emnlp-main.609/.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*, 2023.

Peter H. Westfall and S. Stanley Young. *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. John Wiley & Sons, 1993.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *International Conference on Learning Representations (ICLR)*, 2024.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

Xi Ye and Greg Durrett. The unreliability of explanations in few-shot prompting for textual reasoning. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, 2022.

Kang Min Yoo, Junyeob Kim, Hyuhng Joon Kim, Hyunsoo Cho, Hwiyeol Jo, Sang-Woo Lee, Sang-goo Lee, and Taeuk Kim. Ground-truth labels matter: A deeper look into input-label demonstrations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 2422–2437, 2022.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2309.16042.

Zhanke Zhou, Rong Tao, Jianing Zhu, Yiwen Luo, Zengmao Wang, and Bo Han. Can language models perform robust reasoning in chain-of-thought prompting with noisy rationales? In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, 2024.

Xiangrong Zhu, Yuexiang Xie, Yi Liu, Yaliang Li, and Wei Hu. Knowledge graph-guided retrieval augmented generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8912–8924. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.449.

APPENDIX

Table 2: Appendix evidence guide. Rows group the principal audit tables by the question a reviewer is likely to check; detailed intermediate and negative-result tables remain in their original subsections.

Reviewer question	**Evidence location**	**Main-text claim supported**
Was the comparison constructed as intended?	Tables 1, 3, 4, and 5	Condition definitions, population accounting, leakage checks, and the permuted-control comparison
What does answer-line corruption measure?	Tables 9, 11, 28, and 27	Dominant inconsistency, cross-model ordering, trace-following dominance, and persistence under own reasoning
How does reasoning change the contrast?	Tables 30, 34, 33, and 35	Placebo-normalised growth, signed-mean change, and trace decomposition in Llama
Does the result survive construction choices?	Tables 41, 42, 44, and 45	Dose, length, option-order, and answer-mention sensitivities
Which components carry the cue effects?	Figures 3, 4, and 6; Tables 20, 21, 24, and 56	Shared endpoints, answer-span origin, the writer-output direction, and aligned-only carrier factorisation
How are the endpoints connected?	Tables 23, 54, and 64	Immediate layer-16 MLP reception, negligible selected-carrier query route, and distributed upper-layer mediation
Does it explain generated answers?	Tables 72, 74, 76, 79, and 83	Non-proportional accuracy loss, off-axis errors, component-level tracking nulls, and limited recovery

Appendix roadmap. Appendix A documents construction and audits. Appendices B–D define the readouts and report factorial, polarity, specificity, source, writer-direction, and writer-to-receiver analyses. Appendix E gives the placebo and answer-margin dynamics. Appendices F–I contain robustness, carrier factorisation, localisation, and geometry checks. Appendices J–L give generation, error-channel, component-tracking, cross-dataset, and reproducibility evidence.

A Detailed Data Construction and Audit

A.1 Evaluation funnel

The Llama behavioural and mechanistic analyses use nested populations, summarised in Table 3. Behavioural accuracy is evaluated on all question types. The first-token mechanistic readout requires comparison questions with exactly two candidate answers, a correct Llama clean-baseline generation, and distinct first tokens for the gold and foil. The resulting 1,251-item population is then held fixed for the Qwen mechanistic analyses. Of these items, 1,217 have span-valued answers and 34 have number-valued answers. Eight items are excluded from the trace-conditioned readout because at least one corrupted-arm generation reaches the 384-token limit before emitting an answer line. Their traces are markedly longer than the population median (1,628 versus 360 characters) and end mid-sentence; all eight retain a parseable answer line under the clean baseline. The exclusion is therefore treatment-dependent, although it affects only 0.6% of the mechanistic set. The generation-time intervention makes a further post-treatment restriction to items that become wrong and change their reasoning under plausible corruption, followed by the no-op stability gate.

Table 3: Item-population funnel. Selection through the 1,251-item mechanistic set uses Llama outcomes; Qwen is evaluated on that fixed set.

Stage	n	**Use or exclusion**
Full 2WikiMultiHopQA evaluation	12,576	five-arm Llama behavioural analysis
Comparison questions	3,040	two candidate answers
Mechanistic set	1,251	Llama baseline-correct; fixed for both models
Trace-conditioned set	1,243	eight corrupted generations hit the 384-token limit
Failed-repair set	79	ARM 4A generation incorrect; reasoning changed
Determinism-stable intervention set	76	three no-op-unstable items removed

A.2 Context and exemplar retrieval

Each prompt combines two retrieval products. The evidence context is produced upstream by KG^{2}RAG (Zhu et al., 2025). Starting from semantically retrieved seed chunks, KG^{2}RAG expands the set through knowledge-graph relations and organises the resulting chunks into coherent paragraphs. Our pipeline conditions the model on the resulting `organized_contexts` field. The 2WikiMultiHopQA and HotpotQA exports use the same KG^{2}RAG configuration.

Worked exemplars are retrieved independently from a fixed database shared across models and arms. Its candidate pool contains the 6,040 exact-match-correct chains from an earlier Llama generation pass over the same 12,576-item development split (48.0% exact match). This source pass is distinct from the clean Llama baseline reported in Section 7, which has 6,937 exact-match-correct generations (55.2%); the two item-level EM fields disagree for 2,357 question identifiers. The query is the target question text, encoded using `mixedbread-ai/mxbai-embed-large-v1` through SentenceTransformers (Lee et al., 2024; Reimers & Gurevych, 2019). Database and query embeddings are 1,024-dimensional and L2-normalised, so cosine similarity is computed by their dot product. If a question already has a stored database embedding, that vector is reused; otherwise it is encoded at retrieval time. The three highest-scoring chains are selected. When the query itself belongs to the candidate pool, its score is set to negative infinity before ranking, implementing leave-one-out retrieval. This exclusion is consequential: 1,233 of the 1,251 mechanistic targets are pool members and would otherwise retrieve themselves first. The encoder and retrieval depth are fixed across datasets and are not tuned.

Each retrieved chain is inserted before the target question as an `Example` block containing its question and reasoning. For a given target item, the evidence context, exemplar identities, exemplar order, and source-chain text are fixed across ARM 4A and ARM 4B. Exemplar reasoning is truncated to an 800-character block whose retained-body budget reserves space for the answer line. Among the 3,750 valid mechanistic-set slots, 353 reasoning strings exceed that threshold and 309 of these pair answer strings of different length. The retained reasoning consequently differs across ARM 4A and ARM 4B in 27 slots, each from a different item (0.72% of slots; 2.2% of items). No chain has more than one answer line, so the substitution and truncation rules never select different answer

lines. Excluding the 27 affected items changes the Llama cue-probe ARM 4A−ARM 4B difference from -0.3461 to -0.3514 (95% CI $[-0.4301, -0.2709]$; $n = 1{,}224$). The comparison is therefore near-minimal rather than literally answer-line-only. Neither construction is designed to insert the target item's gold or foil into its exemplars; the stricter audit below identifies the cases in which answer-string containment nevertheless occurs.

A.3 Prompt format and decoding

The archived run configuration recovers the complete prompt. The system message is:

```
You are a precise multi-hop question answerer.  Given a context and a
question, you reason step by step through the intermediate facts in the
context, then give your final answer on a new line starting with
'Answer:'  followed by only the answer itself in as few words as
possible.  Before answering the current question, you may study similar
past questions with their reasoning and answers as guidance.
```

Each retrieved exemplar is formatted as:

```
Example {i}:
Question:  {question}
Reasoning:  {reasoning}
```

The three exemplar blocks precede the target context and question. The user template following those blocks is:

```
{exemplars_block}Now answer the following question.
Context:  {context_str}
Question:  {question}
Reason step by step using only the context above, then end with 'Answer:
X' where X is the answer in as few words as possible.
```

Generation is greedy (`do_sample=False`); no temperature is passed. The model uses float16 inference on CUDA, batch size one, a 6,000-token input-truncation limit, a 384-token generation limit, repetition penalty 1.0, and the end-of-sequence token as the padding fallback. Retrieval depth is three and self-retrieval is leave-one-out.

Exact match is computed by the archived answer parser. Its normaliser lowercases, removes punctuation, strips leading and internal articles, and collapses whitespace. It does not apply Unicode normalisation or strip diacritics, so accented and unaccented spellings remain distinct. Recomputing exact match for all 12,576 clean-baseline generations reproduces the 6,937 stored correct answers exactly, with no unexplained disagreement.

A.4 Automated retrieval-leakage audit

Exemplars are retrieved leave-one-out from an EM-correct pool constructed by an earlier generation pass over the same dataset split. The construction is therefore transductive, and the reported exact-match values measure behaviour within this controlled prompt design rather than standard held-out benchmark performance.

For the Llama 2WikiMultiHopQA mechanistic set, an automated audit recomputes cosine similarities in float64 from the archived embeddings and reconstructs all three exemplars for every target. All 1,251 targets match the archived top-three order within the 6,040-item EM-correct pool, giving 1,251 exact reconstructions and no mismatches. The database's EM flag and candidate-pool flag have identical membership. Across the resulting 3,753 target–exemplar pairs, the audit finds no self-retrieval and no duplicate target questions under either raw or normalised string matching. These checks verify exact and normalised identity only; semantic near-duplicates were not manually adjudicated.

The target's non-binary gold string occurs in an exemplar question or pre-answer reasoning for 78 pairs spanning 73 targets (5.8%). This usually reflects reuse of entities across 2WikiMultiHopQA comparison questions rather than a duplicate question. Eighteen pairs mention both the target gold and foil. Fourteen of these exemplars demonstrate the target's gold entity for a different relation over the same entity pair, two demonstrate the foil, and two are yes/no exemplars without an entity

answer. The pre-specified high-similarity trigger does not fire for any pair: maximum token-set Jaccard similarity is 0.786, while cosine similarity reaches 0.991; the trigger required both cosine similarity at least 0.99 and Jaccard similarity at least 0.80. The threshold was not changed after the results were inspected.

A separate, stricter audit examines the answer lines manipulated by the corruption rather than exemplar questions and pre-answer reasoning. In 18 of the 1,251 mechanistic items, an exemplar answer string under ARM 4A contains the target question's gold or foil answer. This item count measures answer-line containment and is distinct from the 78 target–exemplar pairs across 73 targets reported above.

Of the 18 containment cases, 14 fall in the 467-item changed-trace complete-case comparison. Removing those 14 targets gives a mean-margin difference of $+2.634$ (95% stratified-bootstrap CI $[1.199, 4.051]$; Mann–Whitney $p = 0.0049$), compared with $+2.599$ $[1.184, 4.047]$ and $p = 0.0056$ in the full analysis. The failure-group mean is unchanged to six decimal places because none of the 14 targets is a failed repair. The slightly larger restricted estimate is inconsistent with inflation of the full-sample association by these overlaps.

The audit also characterises answer extraction in the full 12,576-chain exemplar corpus. Terminal answer lines match the gold field for 5,900 chains and the stored prediction field for 4,006; the latter are surface-form variants of an exact-match-correct answer. A further 261 chains (2.1%) match neither field, and 2,409 have no terminal answer line. This audit was not run for Qwen or HotpotQA.

A.5 CORRUPTION CONSTRUCTION

For ARM 4A, a comparison exemplar's plausible replacement is the option in its "A or B" question that is not the gold answer. If no competitor can be parsed, the slot falls back to the unrelated construction. For ARM 4B, the replacement is another chain's gold answer, drawn from the same question-type subset using a deterministic SHA-256 seed derived from the question identifier. Because the manipulation leaves exemplar questions unchanged, the retrieval embeddings and index metadata are copied unchanged from the clean database into both corrupted databases. This construction preserves retrieval identity across arms. An exact contiguous-match audit under both model tokenizers confirms that every one of the 2,292 plausible competitors occurs byte-for-byte in its source exemplar question; no plausible slot requires a token-only or normaliser-only match.

The 1,251 mechanistic items yield 3,753 exemplar slots. Three slots lack an answer line, leaving 3,750 valid slots. Competitor extraction succeeds for 2,292 slots (61.1%), while 1,458 (38.9%) use the unrelated fallback. The independent manifest reconstruction finds 468 prompts with one plausible-corruption slot, 525 with two, and 258 with all three. Hence every mechanistic prompt contains at least one source-linked competitor, and the weighted total $468+2(525)+3(258) = 2{,}292$ reproduces the slot-level count. Consequently, ARM 4A mixes plausible-corruption replacements with unrelated fallbacks rather than forming a pure plausible-corruption condition. The fallback moves the arm towards ARM 4B at the slot level, although we do not assume that prompt-level interactions are exactly linear in the number of source-linked competitors.

The full retrieval database contains a broader set of question types and construction outcomes. Table 4 reports the mutually exclusive stored strategy labels for all 12,576 records. Only `4a_distractor` denotes the plausible-corruption construction. Its 3,113 records form the plausibility-validation population, whereas the 2,292 figure above counts source-linked competitors retrieved by the mechanistic prompts.

A.6 PERMUTED CONTROL

The ARM 4C construction reassigns the already extracted ARM 4A competitors among plausible slots within the same prompt. It builds a directed compatibility graph over those slots and applies only closed cycles of length at least two. Full cycle covers are preferred; otherwise, the construction retains a maximum-cardinality collection of vertex-disjoint cycles and leaves unmatched slots at their ARM 4A values. When several assignments remain, a frozen hash over the canonical sorted list of source–recipient slot pairs selects the lexicographically smallest value. The released manifest records the selected cycles and tie-break inputs. Both tokenizers yield the same 688 eligible items

Table 4: Audit of the ARM 4A construction labels across the full retrieval database. The single no-substitution record is a benign yes/no edge case in which the extracted competitor equals the original answer.

Stored strategy	n	Own?	Construction outcome
`4a_distractor`	3,113	yes	exemplar's own competing option substituted
`4a_fallback_4b`	7,059	no	competing option unavailable; unrelated type-matched answer used
`no_answer_line_kept`	2,403	no	no answer line was available to corrupt
`no_substitution`	1	no	extracted competitor equalled the stored answer

with no blocking failures, so the control extends to Qwen without reconstruction; only the Llama readout was captured in the current environment.

Table 5: Llama permuted control ($n = 688$). The three contrasts satisfy $G^{(4a-4b)} = G^{(4a-4c)} + G^{(4c-4b)}$ item by item. Eligibility requires at least two plausible slots and yields a stronger restricted ARM 4A−ARM 4B gap than the full-set estimate of -0.347.

Arm	Recipient candidate	Present in prompt	Construction
ARM 4A	yes	yes	recipient question's competing answer
ARM 4C	no	yes	same strings permuted across compatible slots
ARM 4B	no	usually no	unrelated same-category answer

Cue-probe contrast	Mean	95% CI
$G^{(4a-4b)}$ on eligible items	-0.507	$[-0.619, -0.396]$
$G^{(4a-4c)}$	-0.585	$[-0.677, -0.495]$
$G^{(4c-4b)}$	$+0.078$	$[-0.004, +0.159]$

Six pairwise screens define compatibility. A reassigned string must not equal the recipient exemplar's gold answer, the target item's gold or foil, or the recipient's existing ARM 4A string under exact, normalised, alias, or token-sequence equality. It must match the recipient's answer form, occur in its source question by exact byte or in-context token sequence, and leave the recipient's truncated reasoning unchanged. These restrictions prevent a surface no-op, insertion of a target-side answer, cross-form substitution, or arm-dependent truncation.

Of the 1,251 mechanistic items, 783 contain at least two plausible slots and are nominally permutable. Ninety-five have no valid cycle after screening, leaving 688 analysed prompts; 638 admit a full cycle cover. This restricted population is not representative of the full mechanistic set. Across the 688 rendered prompts, all blocking invariants pass: ARM 4A and ARM 4C have identical answer-string and answer-token multisets, equal rendered length, byte-identical non-answer text, no arm-dependent truncation, preserved source occurrence, no self-assignment, and no inserted target-question entity. The maximum additive-identity residual is exactly zero. The 4B answer appears elsewhere in its own unchanged prompt for 1.55% of slots, so the ARM 4C−ARM 4B availability contrast has limited rather than zero contamination. Eighteen prompts contain an unpermuted ARM 4A answer that includes the target gold or foil; these slots are byte-identical across ARM 4A and ARM 4C and therefore do not affect their contrast.

All three arms were recaptured together because the original generation environment was not fully recorded. For Llama, the two per-arm mean absolute deviations from the archive are 0.0122 and 0.0135, narrowly exceeding the frozen 0.01 threshold. The analysis nevertheless proceeds under the separately frozen gap-level rule: gap MAE is 0.0190 (threshold 0.02), mean drift is 0.0007 (threshold 0.01), gap correlation is 0.99987, deadbanded sign agreement is 0.9956, and no item has opposite signs. The decision to use the gap-level criteria was recorded before inspecting any ARM 4C contrast. The decomposition $G^{(4a-4b)} = G^{(4a-4c)} + G^{(4c-4b)}$ then holds exactly for every item.

No Qwen ARM 4C contrast is reported. The recorded-environment Qwen extension instead repeats the consistency factorial and cue-position arrival sweep on their frozen manifests (Appendices D and D.3).

ARM 4C controls prompt-wide occurrence, not local placement. Across the 1,551 reassigned strings, the source question precedes the recipient answer line in 43.9% of cases; the median token distance from source occurrence to recipient answer line is 25 (10th–90th percentile: 14–218). These measured positional differences limit the comparison to the original question–answer pairing and local position beyond prompt-wide availability.

A.7 Automated integrity checks

The context-export audit finds 12,576 rows with 12,576 unique question identifiers, no duplicate identifiers, and no null or empty values in the consumed fields. Gold answers are cross-checked against the official 2WikiMultiHopQA development set. The export retains `graph_build_id`, `extraction_id`, `organized_context_hash`, and source-sentence identifiers for provenance.

The exemplar audit verifies the encoder name, vector dimension, L2 normalisation, candidate-pool membership, top-three ranking, and leave-one-out exclusion. It also checks the slot-strategy counts, equality of retrieved exemplar identities and order across ARM 4A and ARM 4B, and the expected truncation-induced reasoning differences. Separate checks reconstruct the population funnel, verify that gold and foil first tokens differ, and reproduce the behavioural exact-match counts. These are gating checks: the analysis scripts stop if a retrieval invariant or an expected population count fails. The 27-item reasoning exception is reported with an exclusion sensitivity rather than treated as an invariant violation.

B Readouts and Statistical Procedures

B.1 First-token margin and readout conditions

For residual state r_i and answer direction $d_i = W_U[g_i] - W_U[f_i]$, the item-level margin is

$$\Delta_i = r_i^\top d_i = \mathrm{logit}(g_i) - \mathrm{logit}(f_i).$$

Here r_i denotes the representation after the cached final-RMSNorm transformation. In implementation, final margins are read directly from output logits, while intermediate residual and component projections use `apply_ln=True`; Appendix I gives the corresponding geometric convention. All mechanistic claims based on Δ_i concern the first-token binary contrast, not the probability of the complete answer string. The cue-probe appends the answer cue immediately. The trace-conditioned readout is implemented by teacher-forcing the reasoning previously generated in the corresponding arm, up to its answer cue, before reading the same margin.

The matched-irrelevant control assigns each of the 1,251 targets one donor trace from a different item, and the same donor trace is supplied to ARM 4A and ARM 4B. Donor and target agree on answer type and question type, and no donor contains the target's gold or foil under the evaluation normaliser. Donor length tracks the target trace closely and without systematic bias: median absolute deviation is 10.6%, mean signed deviation is -0.3%, and 44.5% of donors are longer; the two length distributions have medians of 63 and 64 whitespace tokens. The archived mapping uses 779 distinct donors across all targets, with mean reuse 1.61 and maximum reuse 7, and fully specifies the analysed target–donor pairs. Among the 1,243 complete cases, 775 distinct donors remain; 452 serve one target, while 220, 77, 24, 5, and 1 serve two, three, four, five, and seven targets, respectively. No target is its own donor. Targets sharing a donor are not strictly independent under an item-level bootstrap. The original assignment procedure, tie resolution, and seed were not retained, so we cannot reconstruct the rule that produced the archived mapping. We instead assess its observable consequences: the realised pairs show no excess donor–target question similarity relative to a permuted-donor null, and no single donor cluster materially changes the content-plus-divergence contrast (Appendix E).

Let $G_i^{(r)} = \Delta_{i,4\mathrm{a}}^{(r)} - \Delta_{i,4\mathrm{b}}^{(r)}$ be the plausible-minus-unrelated gap under readout r. For Llama, the cue-to-own interaction decomposes as

$$G_i^{(\mathrm{own})} - G_i^{(\mathrm{cue})} = \big(G_i^{(\mathrm{matched})} - G_i^{(\mathrm{cue})}\big) + \big(G_i^{(\mathrm{common})} - G_i^{(\mathrm{matched})}\big) + \big(G_i^{(\mathrm{own})} - G_i^{(\mathrm{common})}\big),$$

where the steps operationally represent insertion of a matched irrelevant trace, query-relevant content while trace text is fixed across arms, and arm-specific trace divergence. Closure is exact item by item. The last two terms form the content-plus-divergence contrast. Qwen lacks the common-trace readout needed for this three-way decomposition.

B.2 Inference

Confidence intervals for paired margin contrasts, interactions, and head attributions use 10,000 bootstrap resamples over items. Layer-wise plots show pointwise intervals descriptively, while claims about first exclusion use max-t simultaneous bands over the complete trajectory (Westfall & Young, 1993). Two-sided sign tests exclude ties and test whether positive and negative paired differences are equally frequent. Wilcoxon signed-rank tests are reported as secondary rank-sensitive checks. Behavioural exact-match comparisons use paired permutation tests and exact McNemar tests. The generation-time primary comparison uses the pre-specified one-sided exact McNemar test because the recovery direction was fixed before evaluation.

Because matched-trace donors are reused across targets (775 donors for 1,243 complete cases, maximum reuse 7), we also compute donor-clustered bootstrap intervals for the archived two-way contrasts, resampling donors with replacement and carrying each donor's entire target cluster. Clustered intervals are within 5% of the item-level widths in both models. These diagnostics bound the content-plus-divergence estimate's sensitivity to donor reuse; they do not isolate the fixed-trace content mean in the three-way decomposition.

Inferential hierarchy. The Llama consistency factorial first separates the effect shared by corrupted arms from the smaller differences among answer-line constructions. Its three contrasts were pre-specified; the linear scaling analysis is descriptive. Qwen tests replication of the inconsistency interaction and compares the largest arrival step in relative depth, not mechanistic onset, equality of magnitude, or head identity. The factorial-generation rule requires the failure-AUC lower bound to exceed 0.5; the channel split and answer-source audit are post hoc. The polarity-component and writer-source outcome analyses were designed after the main results; their rules were frozen before first access to the real saved arrays, but they remain exploratory. The own-versus-common factorial readout tests whether a shared trace creates persistence. The trace-polarity rule requires both channel effects to resolve negative and calls trace-following dominant only when their clean-scale-relative difference excludes zero. On the 688 eligible items, the primary aligned comparison is ARM 4A versus the permuted control. Cue-position interchange supplies causal evidence for writer and reader influence. Writer heads are selected on 362 aligned items and evaluated on 326 disjoint items; transfer to inconsistency uses 584 held-out factorial items without reselection. The upper-reader transfer uses the same 584 items and no reselection. Specificity rules are evaluated before the post-hoc arm-pattern interpretation. The writer-to-receiver rule calls downstream-query transmission complete above 0.50, partial between 0.25 and 0.50, and weak below 0.25. Query rows hold keys and values fixed and are not a lower bound on total attention-mediated transmission. Fixed-trace interchange separately tests whether upper carriers are suppressed with reasoning present. Logit-lens trajectories, direct attribution, and span summaries are localisation or correlational analyses. We interpret each result at its stated evidence level rather than counting tests on the same items as independent confirmations.

Decision rules for the factorial, polarity, transplant, writer, reader, specificity, receiver-path, and generation analyses were fixed in the analysis code before each full run and were not externally registered; the linear scaling, layer sweeps, error-channel split, and cross-effect comparisons are descriptive or post hoc as stated.

For each readout, the normalised interaction divides the mean plausible-minus-unrelated gap by that readout's mean ARM 4B margin, then differences the two aggregate ratios. The Llama unrelated-arm means are 3.03 at the cue probe and 9.60 after reasoning.[1] Because each denominator is one

[1]The Llama plausible-minus-unrelated cue mean is -0.346 on all 1,251 items and -0.347 on the 1,243 complete cases; the main text rounds both to -0.35.

aggregate quantity rather than a per-item divisor, the measure is not exposed to small per-item denominators. The alternative mean of per-item ratios agrees in sign ($+0.031$) but has a much wider interval; two items have an exactly zero unrelated-arm margin. We use the aggregate ratio only for the own-reasoning versus cue-probe comparison and do not normalise the trace decomposition. Operational patching fractions use a different estimand: numerator and intact-gap denominator are recomputed jointly in every held-out-item bootstrap resample, without clipping. The equivalence analyses are post-hoc sensitivity checks based on 90% intervals, corresponding to two one-sided tests at $\alpha = 0.05$ (Lakens, 2017).

C Plausibility Validation

For each plausible-corruption record in the full retrieval database, we score candidate answer continuations under Llama conditioned on the exemplar question. Scores are mean log-probabilities per answer token rather than total sequence log-probabilities, so longer candidate strings are not penalised merely for containing more tokens. Table 6 reports the three-way comparison for all 3,113 records. The source-linked competitor is substantially closer to gold than the unrelated-corruption answer and is preferred to gold on roughly one third of records.

Table 6: Llama mean per-token answer log-probabilities for all 3,113 plausible-corruption records, conditioned on the exemplar question. The final columns give the fraction of records on which each corruption exceeds gold.

Model	**Gold**	**Plausible**	**Unrelated**	**Plausible > gold**	**Unrelated > gold**
Llama-3.1-8B	-0.35	-0.87	-6.86	33.0%	0.0%

The question-only gold advantage over the source-linked competitor is 0.52 log-probability units. Adding the exemplar's correct reasoning increases it to 1.82 (Table 7). The replacement is therefore plausible from the question alone but becomes less competitive once the exemplar's own correct reasoning is supplied. This manipulation check is limited to exemplar-level plausibility and does not identify the computation underlying query-level answer-margin dynamics.

Table 7: Llama gold-minus-competitor mean per-token log-probability difference over the same 3,113 records. Positive values favour gold.

Model	**Question only**	**Question + correct reasoning**
Llama-3.1-8B	0.52	1.82

Candidate strings can still differ in tokenisation and lexical frequency, which per-token normalisation does not eliminate. The claim supported here is the large ordering between the plausible and unrelated constructions, not equality between plausible and gold candidates.

D Consistency Factorial and Cue-Position Writers

D.1 Reasoning × answer-line factorial

For each eligible plausible-corruption slot, we swap the two option names in the stored reasoning so that it derives the other option from otherwise identical words, then set the answer line independently. The four cells are gold reasoning/gold answer (gg), gold reasoning/foil answer (gf), foil reasoning/foil answer (ff), and foil reasoning/gold answer (fg). Ineligible and fallback slots remain clean in every condition. The builder verifies that gg reproduces ARM 1 and gf reproduces ARM 4A byte for byte on eligible slots. Case- and article-normalised matching yields 2,150 eligible slots across 1,221 items; 142 source-linked slots are excluded because the foil is absent from the reasoning, truncation is unstable, the swap is not invertible, or the options are nested.

Across 1,344 unique exemplar chains, the gold-minus-foil answer log-probability conditioned on the exemplar reasoning is $+2.70$ for the original reasoning and -1.03 for the swapped reasoning.

The swap reverses which option the reasoning supports in 77.4% of chains, so it is a weaker mirror of the original reasoning rather than a perfect semantic inversion.

Table 8: Condition-level cue margins in the Llama consistency factorial ($n = 1{,}221$). Option mass is the log of the summed gold and foil probability.

Cell (reasoning/answer)	Margin (95% CI)	Option argmax	Log option mass
gg (gold/gold)	$+4.156$ $[3.936, 4.381]$	0.962	-0.266
gf (gold/foil)	$+2.970$ $[2.790, 3.155]$	0.952	-0.265
ff (foil/foil)	$+3.806$ $[3.598, 4.016]$	0.951	-0.286
fg (foil/gold)	$+3.100$ $[2.910, 3.295]$	0.968	-0.240

Table 9: Factorial contrasts and pre-specified simple effects. Negative values shift the target cue margin towards the foil.

Role	Contrast	Estimate (95% CI)	Sign p
Factorial	Answer line	-0.240 $[-0.303, -0.180]$	0.0013
Factorial	Reasoning	-0.110 $[-0.156, -0.066]$	2.2×10^{-4}
Factorial	Inconsistency	-0.946 $[-1.015, -0.878]$	9.8×10^{-124}
Simple	gf $-$ gg	-1.187 $[-1.282, -1.092]$	—
Simple	ff $-$ gg	-0.351 $[-0.437, -0.267]$	—
Simple	fg $-$ gg	-1.056 $[-1.139, -0.975]$	—

The plausible-versus-clean effect decomposes as $-1.187 = -0.240 - 0.946$: about 80% is shared inconsistency. Restricting to 803 items on which every manipulated slot flips leaves the inconsistency estimate unchanged at -0.945 $[-1.034, -0.859]$; the answer-line term shrinks to -0.085 $[-0.135, -0.038]$, making it the fragile component. The factorial therefore supports a two-part account, not an output-the-rejected-option rule.

Scaling versus shifting. We regress each non-clean cell's margin on the clean-cell margin and bootstrap paired items 2,000 times. Relative to the unit clean slope, the inconsistency interaction has slope -0.154 $[-0.172, -0.137]$ and intercept -0.306 $[-0.375, -0.235]$. At the mean clean margin of 4.16, these terms contribute -0.64 and -0.31: approximately 68% proportional shrinkage and 32% constant foil-directed shift. This split is descriptive on the logit-margin scale; it does not imply an inconsistency detector. The proportional account does not describe the cross-model ordering: Qwen's clean margin is twice Llama's while its inconsistency effect is one-third, which is the opposite of what scale alone would predict.

Table 10: Linear margin-scale diagnostics relative to the clean gg margin. Sign flips count items moving from a gold-positive clean margin to a foil-negative condition margin.

Condition	Slope (95% CI)	Intercept (95% CI)	Gold→foil flips
gf	0.745 $[0.716, 0.772]$	-0.127 $[-0.222, -0.032]$	4.7%
ff	0.864 $[0.838, 0.888]$	$+0.215$ $[+0.107, +0.326]$	3.8%
fg	0.811 $[0.789, 0.831]$	-0.270 $[-0.362, -0.178]$	5.3%

D.2 Recorded-environment Qwen factorial

We repeat the cue-probe factorial in Qwen2.5-7B-Instruct on the same 1,221 Llama-selected items, construction manifest, and estimator. The inconsistency interaction reproduces with the same sign and 0.297 of Llama's absolute magnitude. The answer-line, reasoning, and coherent-wrong intervals include zero. This does not show that those components are absent: scaling Llama's smaller effects by 0.297 gives predictions inside the observed Qwen intervals. Qwen's clean margin on these items exceeds Llama's, and its clean gold preference is slightly higher (88.0% versus 86.4%), so the attenuation is not attributable to the Llama-based item selection. Normalised to each model's own clean margin, the difference is larger than the absolute comparison indicates.

Table 11: Same-items cue-probe factorial in Qwen and Llama. Intervals are paired-bootstrap 95% intervals over the common 1,221-item manifest.

Quantity	**Qwen2.5-7B**	**Llama-3.1-8B**
Answer line	-0.0553 $[-0.1233, +0.0144]$	-0.2404 $[-0.3034, -0.1801]$
Reasoning	$+0.0216$ $[-0.0340, +0.0781]$	-0.1103 $[-0.1559, -0.0656]$
Inconsistency	-0.2810 $[-0.3345, -0.2297]$	-0.9461 $[-1.0152, -0.8784]$
fg − gg	-0.2594 $[-0.3344, -0.1855]$	-1.0564 $[-1.1389, -0.9751]$
gf − ff	-0.3026 $[-0.3841, -0.2234]$	-0.8359 $[-0.9174, -0.7555]$
ff − gg (coherent wrong)	-0.0337 $[-0.1263, +0.0579]$	-0.3507 $[-0.4370, -0.2671]$
Clean gg margin	$+8.460$ $[+8.041, +8.886]$	$+4.156$ $[+3.936, +4.381]$
Inconsistency / clean margin	-0.0332 $[-0.0394, -0.0273]$	-0.2276 $[-0.2429, -0.2133]$

The resolved replication claim is therefore the inconsistency interaction, not isolation of that interaction or disappearance of the main effects. Qwen's two constituent gaps, fg−gg and gf−ff, are both resolved negative. Its larger clean margin and slightly higher clean gold preference make the attenuation inconsistent with an off-distribution penalty from Llama-based selection; relative to its own clean margin, Qwen is less sensitive still. No Qwen post-reasoning or generation claim is made from this capture.

D.3 Qwen cue-position arrival

We patch the cue-position residual stream entering each of Qwen's 28 layers on 250 selection items. The intact inconsistency gap is -0.2695 $[-0.3790, -0.1701]$, roughly one quarter of Llama's on the corresponding analysis. Every residual interval through layer 17 includes zero; the effect first resolves entering layer 18, reaches 0.700 of the gap by layer 21, and reaches the 1.002 ceiling at layer 27 (Figure 5). The largest arrival step is written by layer 17/28 (0.61 of depth), compared with 15/32 (0.47) in Llama. We compare largest steps rather than first resolution because resolution is power-dependent: Llama's larger gap resolves shares below 0.01 from layer 3 onward. This comparison is not power-matched: the smaller Qwen gap is estimated at the same sample size, and layer-17 attention already resolves before the residual stream does. The result therefore establishes later resolved residual arrival, not later mechanistic onset.

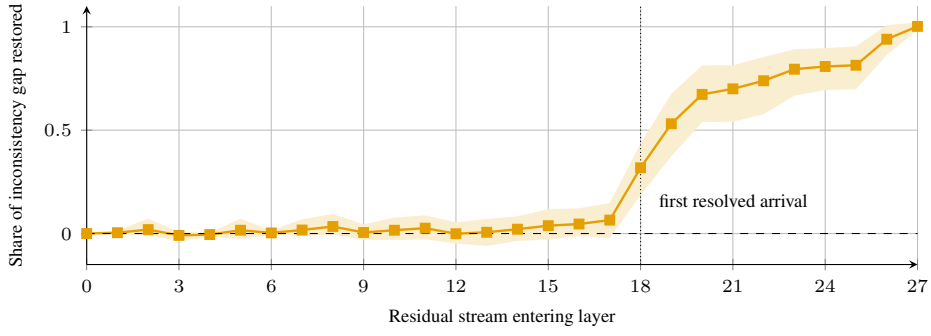

Figure 5: Full-depth residual interchange for Qwen factorial inconsistency ($n = 250$). Shading gives paired-bootstrap 95% intervals. The dotted line marks layer 18, the first residual-stream input whose interval excludes zero. The layer-27 ceiling verifies recovery of the complete effect.

Table 12: Qwen residual and sublayer interventions across the arrival band. Each component is patched independently; shares include downstream influence and are not additive.

Layer	Residual entering layer	Attention output	MLP output
15	0.038 $[-0.029, 0.118]$	—	—
16	0.046 $[-0.013, 0.122]$	0.057 $[-0.029, 0.129]$	0.011 $[-0.078, 0.100]$
17	0.065 $[-0.021, 0.148]$	0.166 $[0.046, 0.265]$	-0.042 $[-0.190, 0.069]$
18	0.318 $[0.186, 0.435]$	0.275 $[0.155, 0.415]$	0.245 $[0.091, 0.462]$
19	0.531 $[0.374, 0.676]$	0.301 $[0.184, 0.461]$	0.023 $[-0.167, 0.157]$
20	0.673 $[0.539, 0.813]$	0.085 $[-0.188, 0.292]$	-0.068 $[-0.395, 0.125]$
21	0.700 $[0.541, 0.814]$	-0.177 $[-0.332, -0.040]$	-0.212 $[-0.542, -0.027]$

Layer-17 attention is the first resolved component, before the residual entering layer 17 resolves. At layer 18, attention and MLP shares are both resolved and have strongly overlapping intervals; the arrival step is therefore not cleanly attributable to attention. At layer 19 only attention resolves, while both layer-21 components oppose the accumulated effect. A pre-specified placeholder test confirms that Llama's layer-15 head indices do not transfer to Qwen layer 18 (share 0.009 $[-0.043, 0.064]$), but head indices are not comparable across models. Two random five-head draws disagree sharply, so Qwen head concentration remains open pending a Qwen-specific ranking.

D.4 Direction transplant

We compute the mean ARM 4A$-$ARM 4C cue-position residual direction on 362 selection items and transplant it bidirectionally on 326 held-out items. Nulls use a norm-matched unrelated-arm placebo direction and two random directions. The pre-specified decision is made at layer 24; the layer sweep is descriptive.

Table 13: Held-out direction transplant. Effects are shares of the intact aligned gap (-0.625 $[-0.761, -0.489]$). Portability is the mean-direction effect divided by the per-item patch effect; coherence is the mean cosine between the shared direction and item differences.

Layer	Item patch	Mean direction	Placebo	Portability	Coherence
8	0.025	0.024	-0.044	0.94	0.80
12	0.088	0.092	-0.051	1.04	0.82
16	0.995	0.492 $[0.406, 0.603]$	0.076 $[-0.023, 0.174]$	0.494 $[0.416, 0.590]$	0.72
20	0.977	0.001 $[-0.047, 0.044]$	0.054	0.001	0.58
24	0.989	-0.082 $[-0.138, -0.041]$	0.071 $[0.018, 0.129]$	-0.083 $[-0.140, -0.041]$	0.54

At layer 16, a common direction carries about half of the per-item effect. From layer 20 upward, it retains moderate geometric coherence but no margin effect; the pre-specified layer-24 decision is therefore item-specific. At layer 24 the same mean direction reproduces 96% $[87, 106]$ of the carriers' natural shift towards the target question and 115% $[102, 132]$ of their shift away from context, while moving the margin slightly in the wrong direction. Coarse span-level attention movement is thus neither sufficient nor the causal quantity identified by the query intervention.

D.5 Layer sweep and sublayer writers

Per-item residual interchange at every layer identifies when the aligned difference becomes available at the cue position. Layer indices denote the residual stream entering each layer. The share rises gradually from layers 7–15 and then jumps from 0.345 entering layer 15 to 0.995 entering layer 16; this is the step written by layer 15. It remains near one thereafter.

Layer-15 attention matches the 0.65 arrival step. Layers 12–14 also exert earlier attention effects, while the layer-13 MLP and layer-16 attention oppose the aligned difference. Placebo and random directions are not uniformly inert, so claims rely on the per-item and held-out head interventions rather than on vector specificity alone.

Table 14: Cue-position arrival sweep. Each layer denotes its input residual stream, and step is the change in recovered share from the preceding layer.

Layer	Share	Step	Layer	Share	Step
0	0.000	0.000	16	0.995	+0.650
1	−0.001	−0.001	17	0.996	+0.001
2	0.000	+0.000	18	1.019	+0.023
3	0.003	+0.004	19	0.991	−0.028
4	0.004	+0.001	20	0.977	−0.015
5	0.003	−0.001	21	0.976	−0.000
6	0.002	−0.001	22	0.985	+0.009
7	0.030	+0.028	23	0.991	+0.006
8	0.025	−0.004	24	0.989	−0.002
9	0.034	+0.008	25	1.000	+0.011
10	0.073	+0.040	26	0.994	−0.006
11	0.098	+0.025	27	0.992	−0.002
12	0.088	−0.010	28	0.989	−0.003
13	0.138	+0.050	29	0.989	+0.000
14	0.299	+0.161	30	0.990	+0.001
15	0.345	+0.045	31	1.008	+0.018

Table 15: Cue-position sublayer interventions around the arrival step. Values are shares of the intact aligned gap. Component effects include direct and downstream influence and need not sum to the residual step.

Layer	Attention output	MLP output	Mean direction	Placebo	Random directions
12	0.162 [0.129, 0.208]	−0.004 [−0.041, 0.030]	0.092	−0.051	+0.058, +0.037
13	0.341 [0.276, 0.420]	−0.140 [−0.220, −0.076]	0.118	0.046	−0.011, −0.009
14	0.278 [0.223, 0.342]	0.011 [−0.030, 0.053]	0.165	0.274	−0.092, +0.003
15	0.661 [0.589, 0.734]	0.261 [0.180, 0.358]	0.135	0.250	+0.032, −0.025
16	−0.134 [−0.206, −0.073]	0.263 [0.211, 0.328]	0.492	0.076	−0.087, −0.207

Held-out head writers. We rank all 32 layer-15 heads by single-head intervention on 362 selection items, then patch nested and control sets on 326 disjoint items. H17, H8, H11, H2, and H1 form the pre-specified top-five set; the selection ranking drops from 0.095 for H1 to 0.041 for H3.

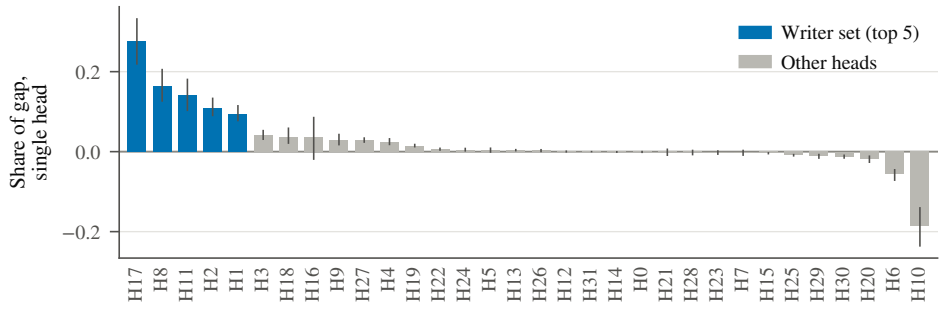

Figure 6: Single-head interchange at layer 15 on the 362-item selection half, as shares of the intact aligned gap with 95% joint-bootstrap intervals. The five-head writer set (blue) was chosen from this ranking and evaluated on the disjoint half (Table 16). H10 and H6 oppose the effect when patched alone; they were not patched jointly.

The top-five/complement interaction is $+0.003$ logits $[-0.002, +0.008]$, so those sets reconstruct the whole-layer effect without resolved interaction. The top-five share of the layer includes one; we therefore claim concentration, not that five heads exceed the layer. The top-ten ratio above one implies a negative net contribution from the remaining 22 heads. On the selection half, H10 and H6 oppose the effect when patched alone (Figure 6), with shares of -0.185 $[-0.238, -0.139]$ and -0.057 $[-0.073, -0.043]$. Because they were not patched jointly and the additivity test covers only the top-five/complement split, this is an observation about the selection ranking rather than a

Table 16: Held-out layer-15 head interventions. Set effects are shares of the intact aligned gap; layer shares divide by the whole-layer effect.

Set	Heads	Share of gap (95% CI)	Share of layer
Whole layer	all 32	0.661 [0.589, 0.734]	1.000
Top 1	H17	0.203 [0.148, 0.255]	0.307 [0.233, 0.370]
Top 3	H17, H8, H11	0.523 [0.445, 0.603]	0.792 [0.677, 0.916]
Top 5	H17, H8, H11, H2, H1	0.724 [0.641, 0.816]	1.096 [0.982, 1.229]
Top 10	top five plus next five	0.895 [0.819, 0.988]	1.355 [1.262, 1.474]
Complement of top 5	other 27	−0.059 [−0.137, 0.015]	−0.090 [−0.218, 0.022]
Random five, draw 1	outside top 10	0.009 [0.000, 0.017]	0.013
Random five, draw 2	outside top 10	0.013 [0.000, 0.025]	0.020

tested opposing component. These cue-position experiments do not locate the condition information before the layer-15 write.

D.6 Transfer of the writers to the inconsistency effect

We form the factorial inconsistency contrast as the mean of fg−gg and gf−ff, patch both donor directions, and average their recovered shares. The evaluation population is the 584 factorial items in the held-out half; the writer heads were selected on the other half using only the aligned plausibility contrast. The two simple gaps are -1.112 $[-1.238, -0.989]$ and -0.821 $[-0.943, -0.696]$, giving a combined gap of -0.966 $[-1.068, -0.862]$. The pre-specified transfer rule calls the writers shared if their share of layer-15 attention has a lower confidence bound above 0.5.

Table 17: Selected checkpoints from the cue-position arrival sweep for inconsistency, alongside the aligned plausibility sweep. Steps are changes from the preceding layer; ranges summarise consecutive layers.

Layer	Inconsistency share (95% CI)	Step	Aligned share
0–6	≈ 0.000	—	0.000–0.004
7	−0.002	−0.002	0.030
10	−0.004	−0.001	0.073
12	−0.016	−0.005	0.088
13	−0.006	+0.009	0.138
14	0.157 [0.135, 0.179]	+0.164	0.299
15	0.202 [0.177, 0.224]	+0.044	0.345
16	0.877 [0.862, 0.891]	+0.675	0.995
17	0.910	+0.033	0.996
18	0.981	+0.071	1.019
19–30	0.97–0.98	≈ 0	0.98–1.00
31	0.998	+0.019	1.008

Table 18: Cue-position sublayer interventions for the inconsistency and aligned plausibility effects. Values are shares of the corresponding intact gap.

Layer	Attention: inconsistency	Attention: aligned	MLP: inconsistency	MLP: aligned
12	−0.003 [−0.009, 0.002]	0.162	−0.004	−0.004
13	0.151 [0.129, 0.171]	0.341	0.032	−0.140
14	0.146 [0.121, 0.169]	0.278	−0.039	0.011
15	0.687 [0.665, 0.711]	0.661	0.054	0.261
16	−0.030 [−0.046, −0.013]	−0.134	0.359 [0.339, 0.381]	0.263

Table 19: Transfer of the layer-15 head sets to factorial inconsistency. The five writers were selected on the aligned contrast, without reselection on these items.

Set	Heads	Share of inconsistency (95% CI)	Share of layer-15 attention
Writers	H17, H8, H11, H2, H1	0.663 $[0.637, 0.689]$	0.965 $[0.925, 1.005]$
Complement	other 27	0.027 $[0.001, 0.055]$	0.040 $[0.001, 0.078]$
Random five, draw 1	H5, H14, H22, H26, H31	0.006 $[0.004, 0.009]$	0.009
Random five, draw 2	H5, H15, H19, H21, H23	0.035 $[0.031, 0.039]$	0.051

The transfer rule is satisfied: without reselection, the writers carry 0.663 of the inconsistency effect and 0.965 of layer-15 attention's share (Table 19). Both effects enter the cue position chiefly in the layer-15 write into the residual stream entering layer 16, but their onset differs: the aligned difference begins accumulating at layer 7, whereas inconsistency is absent there through layer 13. Layer-16 MLP output then carries a further 0.359 of inconsistency. These results identify a shared downstream writer set; they do not show that those heads detect or compute inconsistency.

D.7 Writer source decomposition

We decompose the five writers' differential cue-position output exactly by source span on the 584 held-out factorial items. Reconstructing `hook_z` from the span contributions has maximum error 2.40×10^{-4}.

Table 20: Source decomposition of the writers' differential output for factorial inconsistency. Shares use the complete differential write as the denominator; intervals are paired-bootstrap 95% intervals.

Source span	Share (95% CI)	fg−gg	gf−ff
Manipulated answer	0.841 $[0.832, 0.850]$	0.843	0.839
Template	0.137 $[0.130, 0.144]$	0.133	0.141
Target question	0.019 $[0.017, 0.021]$	0.019	0.019
Context	0.002 $[0.002, 0.003]$	—	—
Manipulated body	0.001 $[0.001, 0.001]$	—	—
Other spans	≈ 0	—	—

Attention mass and differential write localise differently. The template receives 0.60 of the writers' mean attention at the cue but contributes 0.137 of the write. The manipulated answer receives 0.12 in gg and 0.16 in gf yet contributes 0.841. Attention mass alone would therefore identify the wrong dominant source.

D.8 Writer-output direction

For each factorial pair $x \in \{\mathrm{fg} - \mathrm{gg}, \mathrm{gf} - \mathrm{ff}\}$, let

$$\delta_i^{(x)} = \sum_{h \in H_{\mathrm{writer}}} \left(z_{ih}^{(x_1)} - z_{ih}^{(x_2)} \right) W_O^h$$

be the five writers' differential cue-position output, with conditions ordered as named. We estimate $u^{(x)} = \overline{\delta}^{(x)} / \|\overline{\delta}^{(x)}\|$ from 200 items in the selection split, separately for the two pairs, and evaluate it on the 584 held-out factorial items. No evaluated item therefore contributes to its direction. At `blocks.15.hook_resid_mid`, we deliver the projected component $(\delta_i \cdot u)u$, its orthogonal remainder, or a norm-matched random direction at the cue position. Each effect averages the two donor directions as $(\mathrm{recover} - \mathrm{degrade})/2$.

Table 21: Held-out decomposition of the writers' inconsistency effect. Shares jointly resample each intervention with the intact inconsistency gap or writers' total; brackets are paired-bootstrap 95% intervals.

Delivered component	Share of inconsistency gap	Fraction of writers' total
Writers' complete differential output	0.663 $[0.637, 0.689]$	—
Projection $(\delta \cdot u)u$	0.529 $[0.502, 0.557]$	0.798 $[0.762, 0.836]$
Orthogonal remainder	0.138 $[0.111, 0.164]$	0.208 $[0.169, 0.244]$
Norm-matched random direction	0.004 $[-0.013, 0.021]$	0.006 $[-0.020, 0.032]$

The pre-specified rule classifies the write as *single-direction*: the lower confidence bound for the projected fraction exceeds 0.75. Adding the complete δ_i reproduces wholesale writer patching with aggregate relative error 1.43×10^{-3} and median per-item error 5.81×10^{-3} over 40 checks. The pair-specific directions have cosine 0.990. They capture 0.625 and 0.602 of δ_i's mean squared norm, with mean per-item cosines 0.761 and 0.742. Thus the mean direction describes most, but not all, item-level writer variation; the claim is not exact rank one.

Table 22: Geometry and specificity of the writer directions. Answer-axis alignment uses all 1,251 manifest items with distinct answer tokens; clean-output comparisons use the 584 held-out items. References are drawn within the writers' 640-dimensional W_O column space, avoiding the bias of a full-residual-space random vector.

Quantity	fg−gg	gf−ff	Matched reference
Alignment with gold–foil unembedding contrast	-0.0027 $[-0.0039, -0.0015]$	-0.0027 $[-0.0039, -0.0016]$	random $\lvert\cos\rvert$: 0.0120
Cosine with mean clean writer output	0.049	0.066	in-span random: $+0.050/-0.043$
Mean absolute per-item clean-output cosine	0.056	0.065	in-span random: $0.043/0.036$
Top-5 clean-output subspace, uncentred	0.005	0.010	in-span random: $0.007/0.008$
Top-5 clean-output subspace, centred	0.004	0.007	in-span random: $0.008/0.008$

Both pre-specified specificity rules return *specific*: the direction is no closer to the writers' clean-cell output than an in-span random reference and is nearly orthogonal to its dominant modes. Its small negative answer-axis cosine is statistically resolved only because $n = 1{,}251$; at the relevant scale it is indistinguishable from orthogonal, and unembedding the direction yields no coherent lexical structure. Raw and post-final-RMSNorm cosines coincide because the folded layer norm supplies only a scalar rescaling. Together with the direct-logit fraction of -0.017 in Table 23, these measurements rule out a direct linear answer readout from u; they do not rule out information that the layer-16 MLP could decode nonlinearly.

D.9 Writer-to-receiver path interventions

The writers' differential output at the cue, $\delta = \sum_h (z_h(c) - z_h(a))W_O^h$, is delivered to one receiver at a time; every other component reads the unpatched stream. For a query receiver, we recompute each listed head's cue-position query from its own normalisation of the recipient residual plus δ using W_Q and b_Q. Reconstruction agrees with the model's unmodified query to within 2% relative error (observed maximum 4.41×10^{-4}). Keys and values are not modified at any position, including the cue position's own key and value. The query rows therefore measure the query route specifically, not every attention-mediated route. Paths are nonlinear and non-additive, so the rows are not expected to sum to the writers' total.

Table 23: Receiver-specific paths for the five writers' inconsistency effect ($n = 584$). Shares are given relative to the intact inconsistency gap and to the writers' total intervention effect; intervals are joint-bootstrap 95% intervals.

Receiver	Share of inconsistency gap	Fraction of writers' total
Five carrier queries	$0.002\ [-0.001, 0.005]$	$0.002\ [-0.002, 0.007]$
Thirty selected-head queries	$0.006\ [0.002, 0.010]$	$0.010\ [0.003, 0.016]$
All layer-24–31 queries	$0.042\ [0.034, 0.050]$	$0.063\ [0.052, 0.074]$
All layer-16–31 queries	$0.297\ [0.267, 0.327]$	$0.448\ [0.409, 0.484]$
Layer-16 MLP	$0.262\ [0.246, 0.280]$	$0.396\ [0.370, 0.423]$
All layer-16–31 MLPs	$0.304\ [0.281, 0.327]$	$0.458\ [0.426, 0.492]$
Direct logits	$-0.011\ [-0.018, -0.005]$	$-0.017\ [-0.027, -0.008]$
Writers' total	$0.663\ [0.637, 0.689]$	—

The pre-specified rule calls the downstream-query result partial because 0.448 lies between its 0.25 and 0.50 thresholds. The selected carriers, thirty-head band, and all layer-24–31 queries receive little of the writers' effect through their queries. The layer-16 MLP is the largest single immediate receiver measured here; how its output later reaches the upper-layer readers remains unresolved.

D.10 Transfer of the upper-layer readers to inconsistency

We apply the aligned reader sets to the same 584 held-out factorial items without reselection. The inconsistency contrast is the mean of fg−gg and gf−ff, evaluated bidirectionally at the cue position, with an intact gap of $-0.968\ [-1.070, -0.863]$. The pre-specified rule first requires the all-upper interval to exclude zero, then calls the five carriers shared when their lower bound is at least 0.069, half the aligned plausibility estimate of 0.138. The decision is shared.

Table 24: Transfer of frozen cue-position component sets from aligned answer plausibility to factorial inconsistency. Shares use each contrast's own intact gap. Aligned intervals were recomputed from the archived scope run with the same joint bootstrap. The two contrasts use different evaluation populations, so their similarity is descriptive rather than a paired test.

Intervention	Aligned share (95% CI)	Inconsistency share (95% CI)
Layer-15 attention	$0.661\ [0.589, 0.734]$	$0.687\ [0.665, 0.711]$
Five layer-15 writers	$0.724\ [0.641, 0.816]$	$0.663\ [0.637, 0.689]$
Residual entering layer 24	$0.998\ [0.978, 1.019]$	$0.972\ [0.966, 0.977]$
All attention, layers 24–31	$0.695\ [0.640, 0.756]$	$0.687\ [0.662, 0.713]$
Thirty attribution-selected upper heads	$0.278\ [0.240, 0.319]$	$0.217\ [0.195, 0.240]$
Five upper-layer carriers	$0.138\ [0.109, 0.171]$	$0.116\ [0.097, 0.135]$
Random upper five, draw 1	—	$-0.021\ [-0.026, -0.017]$
Random upper five, draw 2	—	$0.002\ [0.001, 0.004]$

The all-upper intervention carries 0.687 of inconsistency, and the five carriers carry 0.116, or 0.168 $[0.142, 0.196]$ of the all-upper effect. The two constituent factorial gaps agree: the carrier shares are 0.110 for fg−gg and 0.123 for gf−ff, while the all-upper shares are 0.687 and 0.688. The corresponding thirty-head shares are 0.216 and 0.219. The layer-24 residual ceiling reproduces 0.972 overall. Item by item, this row differs from the same intervention in the earlier sweep, run in a separate session, by at most 0.113 logits with no systematic offset; we record this as the cross-session reproducibility tolerance for half-precision patching.

The same selected output groups therefore carry both effects, but this does not establish that they encode the same signal. The carrier intervals overlap and the thirty-head intervals meet at 0.240, but there is no interval for either cross-effect difference. The descriptive point-estimate differences therefore do not establish weaker mediation for inconsistency. Query factorisation, direction transplant, and the span-level attention signature remain tested only on the aligned contrast.

D.11 SPECIFICITY CONTROLS

We apply frozen component sets to three distinct perturbations on the same 600 items: removing exemplar reasoning while retaining answers (ARM 2), blanking exemplar answers (ARM 3), and corrupting the answer line (ARM 4A). Each entry divides the bidirectional recovery effect by that perturbation's own clean-relative gap. The pre-specified classification rule returns *mixed* for the writers and *mixed* for the carriers; this verdict precedes the post-hoc pattern interpretation below.

Table 25: Specificity controls on the 600-item evaluation set. The first row gives each perturbation's clean-relative margin gap; remaining rows are recovered shares with joint-bootstrap 95% intervals.

Component set	ARM 2: reasoning removed	ARM 3: answers blanked	ARM 4A: answer corrupted
Perturbation gap	-1.111 $[-1.275, -0.945]$	-0.731 $[-0.835, -0.629]$	-1.519 $[-1.678, -1.366]$
Five writers	-0.137 $[-0.248, -0.051]$	0.418 $[0.361, 0.475]$	0.594 $[0.562, 0.627]$
All layer-15 attention	0.208 $[0.159, 0.253]$	0.498 $[0.446, 0.557]$	0.635 $[0.609, 0.661]$
Random layer-15 heads	0.032 $[0.025, 0.040]$	0.007 $[-0.001, 0.014]$	0.004 $[0.002, 0.007]$
Five carriers	0.079 $[0.058, 0.102]$	0.016 $[-0.012, 0.040]$	0.090 $[0.073, 0.108]$
All layer-24–31 attention	0.424 $[0.370, 0.470]$	0.474 $[0.426, 0.516]$	0.652 $[0.626, 0.678]$
Residual entering layer 24	1.025 $[0.988, 1.065]$	0.970 $[0.942, 1.001]$	0.952 $[0.942, 0.961]$

Post hoc, the writers carry the two perturbations that modify exemplar answers but move slightly opposite to recovery when reasoning bodies are removed and answers are retained. The independent answer-span decomposition in Table 20 supports the resulting hypothesis that they read answer spans rather than act as generic margin movers. A fourth control that reverses context passages produces an unresolved margin perturbation of -0.041 $[-0.255, +0.171]$; its component shares are therefore uninterpretable and are not reported.

D.12 FACTORIAL AFTER A COMMON REASONING TRACE

We read the four factorial cells in one process at the answer cue and after a common trace. The trace is the clean prompt's own generated reasoning, cut after its final `Answer:` and appended identically to every cell, as for the other common-trace readouts. The pre-specified rule classifies the inconsistency contrast as persisting if it is resolved negative after the trace and the retained fraction (trace mean over cue mean, jointly resampled) has a lower bound above 0.5. The cue readout reproduces Table 8 exactly, and the clean cell after the trace matches the archived teacher-forced clean margins (MAE 0.015, $r = 0.99999$, 1,214 items).

Table 26: Factorial contrasts at the answer cue and after one common teacher-forced trace ($n =$ 1,221). Relative values divide each contrast by the mean clean (gg) margin at the same readout: 4.16 at the cue and 13.86 after the trace.

Contrast	Cue	After trace (95% CI)	Retained fraction	Relative (cue)	Relative (trace)
Answer line	-0.240	-0.130 $[-0.191, -0.071]$	0.54 $[0.33, 0.75]$	-0.06	-0.01
Reasoning	-0.110	-0.100 $[-0.136, -0.063]$	0.91 $[0.55, 1.55]$	-0.03	-0.01
Inconsistency	-0.946	-3.047 $[-3.190, -2.908]$	3.22 $[3.02, 3.44]$	-0.23	-0.22
gf − gg	-1.187	-3.177	2.68	-0.29	-0.23
ff − gg	-0.351	-0.230	0.65	-0.08	-0.02
fg − gg	-1.056	-3.147	2.98	-0.25	-0.23

The raw inconsistency effect triples because every margin expands under the trace; relative to the clean margin it is unchanged. The answer-string effects shrink relative to that margin. The linear scaling fit gives an inconsistency slope of -0.146 $[-0.183, -0.108]$ and intercept -1.021 $[-1.537, -0.524]$ after the trace, preserving the two-thirds proportional, one-third constant split observed at the cue. The answer-line term here is not the same estimand as ARM 4A−ARM 4C, whose raw gap is stable under a common trace (-0.585 to -0.616).

D.13 Own-trace and in-policy factorial readouts

We also read each factorial cell after its own generated reasoning. Own traces are complete in every cell for 1,209 of 1,221 items. The final column compares each own-trace contrast with the corresponding common-trace contrast on matched complete cases.

Table 27: Factorial effects after each cell's own generated trace. Relative values divide by the own-trace clean margin; differences are own minus common trace.

Contrast	Own trace (95% CI)	Relative to clean	Own−common (95% CI)
Answer line	-0.168 $[-0.251, -0.084]$	-0.012	-0.038 $[-0.097, +0.020]$
Reasoning	-0.237 $[-0.334, -0.142]$	-0.017	-0.141 $[-0.228, -0.054]$
Inconsistency	-3.056 $[-3.212, -2.906]$	-0.221	-0.009 $[-0.069, +0.053]$
Plausible versus clean	-3.224 $[-3.394, -3.061]$	-0.233	-0.046 $[-0.107, +0.017]$
Coherent wrong versus clean	-0.405 $[-0.540, -0.272]$	-0.029	-0.178 $[-0.300, -0.063]$
Correct answer, disagreeing reasoning	-3.293 $[-3.469, -3.121]$	-0.238	-0.149 $[-0.244, -0.053]$

The inconsistency effect is unchanged when each condition reasons for itself: the own-minus-common estimate is -0.009 and its interval spans zero. The common teacher-forced trace therefore does not manufacture the result. At the in-policy position that emits the first reasoning token, before any answer cue is appended, the inconsistency projection is unresolved at every layer and remains $+0.006$ $[-0.003, +0.015]$ in the final residual stream. The answer-line projection reaches only -0.012 $[-0.022, -0.001]$ there. The factorial effects analysed in this paper are properties of the answer-cue readout, not established shifts at reasoning onset.

D.14 Trace-polarity decomposition

We swap the polarity of the common trace to separate polarity-invariant evidence (e) from trace-following (b). Of 1,221 factorial items, 1,166 pass the frozen gates; 37 are excluded because an option is absent from the trace, 14 because the swap is not invertible, and four because the options are nested. The manipulation check passes: under the polarity-swapped clean trace, the mean margin is -4.725 $[-5.039, -4.405]$, and 80.0% of items are negative. The clean channel scales are $e_{gg} = +4.573$ $[+4.380, +4.769]$ and $b_{gg} = +9.298$ $[+9.116, +9.484]$.

Table 28: Trace-polarity decomposition on 1,166 gated factorial items. The final column is the difference between the trace-following and evidence effects after dividing each by its own clean-channel scale.

Contrast	Evidence e (95% CI)	Trace-following b (95% CI)	Relative $b - e$ (95% CI)
Answer line	-0.291 $[-0.358, -0.229]$	$+0.160$ $[+0.118, +0.203]$	$+0.081$ $[+0.065, +0.097]$
Reasoning	-0.110 $[-0.143, -0.078]$	$+0.003$ $[-0.027, +0.034]$	$+0.024$ $[+0.016, +0.033]$
Inconsistency	-0.582 $[-0.644, -0.519]$	-2.507 $[-2.632, -2.386]$	-0.142 $[-0.157, -0.128]$
Plausible versus clean	-0.873 $[-0.963, -0.785]$	-2.347 $[-2.475, -2.220]$	-0.061 $[-0.080, -0.043]$
Coherent wrong versus clean	-0.401 $[-0.478, -0.329]$	$+0.164$ $[+0.115, +0.213]$	$+0.105$ $[+0.087, +0.124]$
Correct answer, disagreeing reasoning	-0.692 $[-0.762, -0.623]$	-2.504 $[-2.633, -2.379]$	-0.118 $[-0.134, -0.102]$

The pre-specified rule returns *both channels resolved, trace-following dominant*. Relative to their respective clean scales, inconsistency is -0.127 $[-0.140, -0.114]$ in evidence and -0.270 $[-0.281, -0.258]$ in trace-following. A coherent wrong exemplar leaves trace-following intact, whereas disagreement between the exemplar's answer and reasoning sharply reduces it. This identifies where the inconsistency effect lies; it is not a claim that the model becomes unfaithful.

E Supplementary Answer-Margin Dynamics

Table 33 gives the inferential details on one common complete-case population. The fixed-trace content step has a non-uniform sign distribution but a mean interval spanning zero; neither result substitutes for the other. This decomposition is Llama-only.

E.1 UNRELATED-ARM PLACEBO AND CUE-TO-TRACE DIRECTION

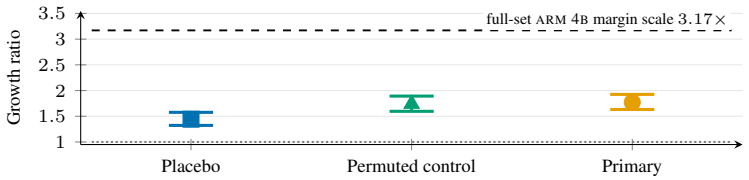

Figure 7: Growth relative to the cue probe under one common clean trace ($n = 684$). Points are later mean $|G|$ divided by cue mean $|G|$, with paired-bootstrap 95% intervals; the dotted line marks no growth. The corruption-bearing contrasts exceed the placebo by 21–23%. For scale only, the dashed 3.17× ARM 4B reference uses the full trace-complete set ($n = 1{,}243$), not the paired subset. Colour and marker shape redundantly identify contrast.

To measure how much raw absolute-gap growth arises without a semantic contrast, we construct ARM 4B′ by permuting ARM 4B's replacement records among chains of the same answer form. The replacement multiset is unchanged, no chain receives its own source record, and every prompt changes in at least one slot; duplicate answer strings leave 8.8% of slots textually unchanged, so 91.2% differ. This arm is a later-readout placebo rather than another corruption condition: both sides use the same unrelated-answer construction. Its later value is read under one common clean-arm trace, so the two arms receive identical reasoning text. On all 1,251 items, cue and common-trace magnitudes are 0.836 (95% CI $[0.793, 0.884]$) and 1.134 ($[1.080, 1.189]$), giving growth $+0.298$ ($[+0.231, +0.363]$) and a ratio of means of 1.356 ($[1.267, 1.449]$).

Table 29: Descriptive raw absolute-gap growth and cue-to-trace association across Llama contrasts. The placebo and the two 688-item rows use one common clean-arm trace; the remaining rows use the named trace condition. "Implied" is the sign agreement $0.5 + \arcsin(\rho)/\pi$ under a bivariate-normal reference. Observed and implied agreement differ by at most approximately 0.05, so agreement is not treated as independent evidence beyond rank correlation.

Contrast / trace-conditioned readout	n	$\|G\|$ cue	$\|G\|$ later	Growth	Spearman ρ	Sign agree.	Implied
Placebo, ARM 4B vs. ARM 4B′	1,251	0.836	1.134	+0.298	0.218	0.584	0.570
$G^{(4c-4b)}$, common trace	688	0.790	1.247	+0.457	0.411	0.686	0.635
$G^{(4a-4b)}$, common trace	688	1.167	2.060	+0.894	0.427	0.646	0.640
$G^{(4a-4b)}$, matched irrelevant trace	1,243	1.069	1.176	+0.107	0.385	0.628	0.626
$G^{(4a-4b)}$, own trace	1,243	1.070	2.101	+1.031	0.396	0.639	0.630

For an inferential comparison, we intersect the placebo, cue, and common-clean-trace captures, leaving 684 of the 688 permuted-control items. Table 30 reports ratios of means, not the heavy-tailed mean of per-item ratios. Both corruption-bearing contrasts exceed the placebo by 21–23%, and both excess intervals exclude one. The permuted-control result shows that this excess does not depend on the unrelated arm's prompt-wide answer-string availability. Ratios of the raw changes are much larger (2.453 and 2.076) and are reported beside the placebo-normalised estimate rather than substituted for it.

Table 30: Paired comparison at the common-clean-trace readout ($n = 684$). Growth is later mean $|G|$ divided by cue mean $|G|$; excess divides each corruption-bearing growth ratio by the placebo ratio within each bootstrap resample. The raw-change ratios expose scale dependence and are not the primary estimand.

Contrast	Cue $\|G\|$	Common-trace $\|G\|$	Growth ratio (95% CI)	Excess / placebo	Raw-change ratio
Placebo, ARM 4B vs. ARM 4B′	0.827	1.194	1.443 [1.322, 1.576]	—	—
$G^{(4a-4b)}$	1.165	2.065	1.772 [1.632, 1.925]	1.228 [1.095, 1.373]	2.453
$G^{(4a-4c)}$	1.029	1.791	1.740 [1.596, 1.892]	1.205 [1.073, 1.351]	2.076

Growth ratios depend strongly on cue magnitude because binning on a denominator induces regression to the mean. To test whether the pooled excess is driven by the placebo's smaller cue baseline

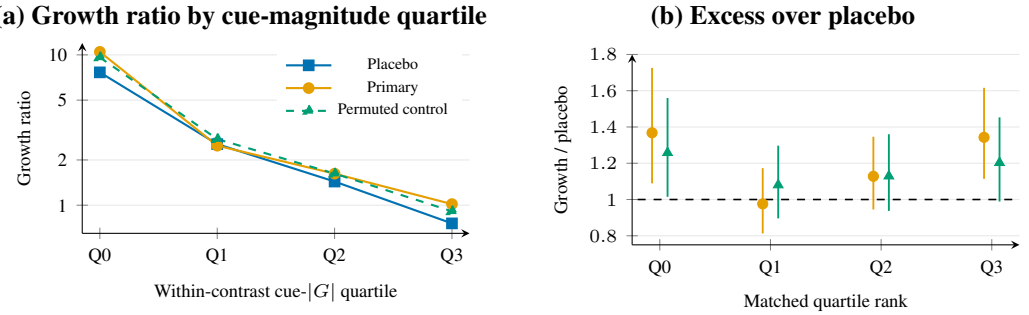

Figure 8: Quartile sensitivity on the paired 684-item population. Panel (a) shows the strong denominator dependence of all three raw growth ratios. Panel (b) shows the two corruption-bearing ratios divided by the placebo within matched quartile ranks, with paired-bootstrap 95% intervals. Colours and markers match Figure 7; the exact values are retained in Table 31.

without selecting one contrast on the other's scale, we divide each contrast into quartiles of its own cue $|G|$ distribution and form the excess within matched quartile ranks (Figure 8; Table 31). The raw ratios vary by roughly an order of magnitude in every contrast, but their quotient varies by at most $1.4\times$ and its four point estimates are non-monotone. For ARM 4A−ARM 4C, every quartile interval covers the pooled excess; for ARM 4A−ARM 4B, three of four do, with the exception in an interior quartile. Because growth falls steeply as cue $|G|$ rises across strata, the approximately 40% larger cue magnitudes of the corruption-bearing contrasts would tend to depress, rather than inflate, their growth ratios relative to the placebo. This directional argument is inferred from the between-stratum trend, not from an estimated within-stratum slope. The analysis supports the pooled result as more than a cue-baseline artefact, but the wide intervals do not establish either homogeneity or an excess separately in each quartile.

Table 31: Cue-magnitude sensitivity on the paired 684-item population (approximately 171 items per stratum). Each contrast is stratified by its own cue-$|G|$ quantiles before excesses are formed within matched quartile ranks, avoiding direct selection on only one denominator. Quantile matching does not equalise absolute cue magnitudes: the corruption-bearing contrasts remain roughly 40% larger than the placebo within corresponding strata.

Cue-$\lvert G\rvert$ quartile	Placebo ratio	ARM 4A−ARM 4B ratio	Excess (95% CI)	ARM 4A−ARM 4C ratio	Excess (95% CI)
Q0	7.653	10.472	1.368 [1.089, 1.726]	9.625	1.258 [1.015, 1.560]
Q1	2.550	2.489	0.976 [0.813, 1.173]	2.754	1.080 [0.896, 1.297]
Q2	1.437	1.621	1.128 [0.945, 1.346]	1.623	1.129 [0.937, 1.360]
Q3	0.757	1.017	1.343 [1.114, 1.616]	0.911	1.203 [0.989, 1.453]
Pooled	1.443	1.772	1.228 [1.095, 1.373]	1.740	1.205 [1.073, 1.351]

The placebo begins near the primary contrast's cue magnitude, so much item-level separation exists without a semantic contrast. Growth is present with one, two, and three plausible slots ($+0.875$, $+1.114$, and $+1.146$), but we do not infer a dose trend because the subgroups are compositionally different and lack trend intervals. Supplying reasoning unrelated to the target question yields only $+0.107$. The robust conclusion is therefore a modest relevant-trace excess above a large placebo floor, not generic trace insertion or a graded continuum of corruptions.

The raw arm margins grow by about $3.2\times$, faster than any contrast in Table 29: 1.356 for the full placebo, 1.740 for aligned ARM 4A−ARM 4C, 1.772 for common-trace ARM 4A−ARM 4B, 1.96 for own-trace ARM 4A−ARM 4B, and 1.10 for the matched-irrelevant trace. Every contrast therefore shrinks relative to the readout's arm-margin scale. We report raw growth because it captures redistribution in logit units, while avoiding the stronger claim of scale-invariant amplification. Values from different populations are not divided: Table 30 performs the placebo comparison on one paired set.

Two diagnostics bound dependence on the archived donor assignment. Removing each donor's complete target cluster shifts the Llama content-plus-divergence estimate by at most 0.022 without changing its sign. A separate permutation test compares target–donor question similarity with permutations of the same donor multiset. Across 1,243 pairs, the observed mean cosine is 0.3325,

Table 32: Llama-3.1-8B correct-answer margins before and after reasoning. At the cue probe, plausible corruption lowers the gold-minus-foil margin more than unrelated corruption; the trace-conditioned signed-mean difference is smaller. The plausible–unrelated signed-mean change is tested by the paired interaction in Section 6. Paired bootstrap; $n = 1{,}251$ before reasoning and $n = 1{,}243$ after reasoning.

	Correct-answer margin Δ (mean, 95% CI)	
Arm	**Cue-probe (before reasoning)**	**Trace-conditioned (after reasoning)**
ARM 1 clean	4.14 [3.93, 4.36]	13.85 [13.65, 14.05]
ARM 4A plausible corruption	2.67 [2.51, 2.84]	9.53 [9.31, 9.75]
ARM 4B unrelated corruption	3.02 [2.85, 3.19]	9.60 [9.39, 9.80]

Table 33: Llama trace-readout decomposition on the 1,243 complete cases. The three steps sum to the total item by item before rounding. Percentages and sign tests use non-tied differences.

Model	**Contrast**	**Mean D (95% CI)**	**% positive**	**Sign p**	**Wilcoxon p**
Llama-3.1-8B	Cue → own reasoning	$0.2777\ [0.113, 0.447]$	57.2	4.9×10^{-7}	2×10^{-6}
	Cue → matched (irrelevant insertion)	$0.0985\ [0.012, 0.184]$	52.3	0.112	—
	Matched → common (fixed-trace content)	$0.105\ [-0.039, 0.247]$	55.1	3.9×10^{-4}	—
	Common → own (trace divergence)	$0.074\ [-0.030, 0.183]$	49.2	0.613	—
	Matched → own (content + divergence)	$0.1791\ [-0.002, 0.362]$	54.9	5.9×10^{-4}	3.3×10^{-3}

compared with a null mean of 0.3320 (SD 0.0015; upper-tail $p = 0.38$). The realised assignment is therefore not detectably closer on this embedding measure than an arbitrary pairing of the same donors.

For the own-reasoning versus cue-probe comparison, dividing each readout's mean gap by its mean ARM 4B margin yields a Llama interaction of $+0.107$ (95% CI $[+0.082, +0.132]$). The plausible-minus-unrelated gap decreases from 11.4% of the unrelated-arm mean margin at the cue probe to 0.7% after reasoning. These aggregate ratios are not used for the matched-reasoning decomposition; Appendix B defines the estimand and reports the per-item-ratio sensitivity.

Table 35 tests a trace-noise alternative within the plausible–unrelated contrast. The common-clean-trace condition retains about 73% of the own-trace signed change, with nearly unchanged rank correlation. Observed sign agreement closely matches the value implied by ρ and is reported as a diagnostic rather than independent evidence. Arm-specific trace divergence therefore contributes to the own-trace signed change but is not sufficient to explain it. Table 30 supplies the separately reconciled common-trace magnitude comparison.

E.2 Which changes survive the permuted control?

We recapture ARM 4A, ARM 4B, and ARM 4C together on the 688 permuted-control items at the cue probe and after appending the same clean-arm reasoning trace to every prompt. The two contrasts are therefore compared within item, position, and process, with arm-specific trace divergence fixed at zero.

Across the same transition, the ARM 4A, ARM 4C, and ARM 4B margins change from 2.539 to 9.407, 3.124 to 10.023, and 3.046 to 9.642. The first two gains are nearly identical (6.87 and 6.90), whereas the unrelated-arm gain is 6.60. The full-set common-trace plausible–unrelated change is $+0.204$, consistent with $+0.272$ on this subset, so population selection does not explain the contrast difference. The signed-mean result is therefore specific to contrasts involving the unrelated arm.

E.3 Plausible–unrelated signed mean, absolute magnitude, and direction

For $G^{(4a-4b)}$, the interaction D compares signed means and must be distinguished from the item-level gap magnitude. In Llama, the signed mean moves from -0.347 at the cue probe to -0.069 after the model's own trace, whereas the mean absolute gap increases from 1.070 to 2.101 (change $+1.031$, 95% CI $[+0.899, +1.167]$). The gap as a share of the mean arm margin falls from 0.551 to 0.280 because the arm margins increase substantially. The signed mean therefore moves towards

Table 34: Llama paired ARM 4A−ARM 4B differences before and after reasoning. Confidence intervals use 10,000 paired bootstrap resamples; percentages and exact two-sided sign tests use non-tied items.

Model	Readout	Mean ARM 4A − ARM 4B (95% CI)	Median	% neg.	Sign p	n
Llama-3.1-8B	Cue-probe	-0.35 $[-0.43, -0.27]$	-0.28	61.4	8.2×10^{-16}	1,251
	Trace-conditioned	-0.07 $[-0.24, +0.11]$	$+0.12$	47.9	0.16	1,243

Table 35: Llama plausible–unrelated cue-, own-trace-, and common-clean-trace directional readouts on the same 1,243 complete cases. The common condition supplies the clean-arm trace to both prompts, eliminating arm-specific trace divergence. Changes are relative to the cue probe; brackets give paired-bootstrap 95% intervals. Absolute-gap growth is compared separately on the fully paired 684-item population in Table 30; the unreconciled cross-population magnitude summary is omitted.

Quantity	Cue probe	Own trace	Common trace
Signed mean	-0.347	-0.069	-0.143
Signed change from cue	—	$+0.278$ $[+0.113, +0.447]$	$+0.204$ $[+0.075, +0.328]$
Spearman ρ with cue	—	0.396	0.419
Observed / implied sign agreement	—	$0.639/0.630$	$0.645/0.638$
Shrinkage R^2	—	0.116	0.178

zero while the typical absolute item-level difference grows; Table 36 shows that this mean change does not extend to $G^{(4a-4c)}$.

A shrinkage account would predict that the trace-conditioned gap is approximately a scaled copy of the cue-probe gap. In Llama, a least-squares fit through the origin explains little item-level variance ($R^2 = 0.116$); the fitted scale is 0.732, unlike the 0.200 implied by the signed-mean ratio. Rank correlation is 0.396. With a ± 0.05 deadband, the sign-concordance accounting is:

The exact two-sided binomial test against chance gives $p = 8 \times 10^{-22}$. This restates the dependence already captured by ρ: observed agreement (0.639) is close to the bivariate-normal implication (0.630). The gap is therefore neither preserved nor randomly reassigned: its direction remains related across readouts, but more than one third of items reverse sign. The permuted-control contrast exhibits comparably weak association without a signed-mean change.

E.4 Position-resolved trajectory

To locate when direction and magnitude change, we append the canonical answer cue after tokenised prefixes at 0, 25, 50, 75, and 100% of each trace. We evaluate both the per-arm traces and a common trace supplied to both arms. In the per-arm analysis, 60 of 1,242 endpoint-comparable items exceed the frozen 0.1-logit endpoint-agreement tolerance, primarily among longer traces, and are excluded from the trajectory; retaining them changes every mean by less than 0.03 and sign agreement by less than 0.001.

Directional agreement falls to approximately 0.63 within the first quarter and changes little thereafter, while the mean absolute gap grows across every subsequent prefix. The common-trace condition shows the same early loss of directional agreement (1.000, 0.633, 0.647, 0.659, 0.641) and later magnitude growth (1.069, 0.861, 1.265, 1.424, 1.827). Its initial magnitude dip is specific to the shared-trace construction. The signed mean is non-monotone at the interior points, so we interpret the temporal separation between early decorrelation and accumulating magnitude rather than individual fluctuations.

E.5 Layer-band decomposition across readouts

Cached layer-level direct logit attribution permits a correlational decomposition of the Llama gap into the candidate attention layers and the remainder. Item-level closure is exact.

The candidate band's absolute contribution grows, but the remainder grows faster; direct attribution is correlational and does not by itself identify a head-level causal account. Post-hoc equivalence is

Table 36: Same-process common-trace comparison on the permuted-control Llama population. Magnitude growth and weak cue-to-trace association hold for both contrasts; the signed mean moves towards zero only in the contrast involving ARM 4B. Intervals are paired-bootstrap 95% intervals.

Quantity	**Population / transition**	$G^{(4a-4c)}$	$G^{(4a-4b)}$
Signed mean, cue probe	$n = 688$	$-0.585\ [-0.677, -0.495]$	$-0.507\ [-0.619, -0.396]$
Signed mean, common trace	$n = 688$	$-0.616\ [-0.788, -0.444]$	$-0.235\ [-0.440, -0.038]$
Signed change	cue $\rightarrow$ common	$-0.031\ [-0.194, +0.132]$	$+0.272\ [+0.089, +0.453]$
Mean absolute gap	cue $\rightarrow$ common	$1.025 \rightarrow 1.787$	$1.167 \rightarrow 2.060$
Absolute change	cue $\rightarrow$ common	$+0.762\ [+0.636, +0.888]$	$+0.894\ [+0.750, +1.033]$
Spearman ρ with cue	$n = 688$	0.411	0.427
Observed / implied sign agreement	$n = 688$	$0.686/0.635$	$0.646/0.640$
Signed change, two slots	$n = 492$	-0.018	$+0.211$
Signed change, three slots	$n = 196$	-0.062	$+0.423$

Table 37: Directional accounting for the Llama plausible-versus-unrelated gap between cue-probe and own-trace readouts on the 1,243 complete cases. "Live" items exclude cases deadbanded at either readout.

Cue-to-trace sign relation	**Llama**
Same sign	760
Opposite signs	430
Both deadbanded	1
One deadbanded	52
Agreement among live items	0.639

not established at 20%, 25%, 33%, or 50% of the cue-probe effect. The tightest supported absolute bound is 0.216, or 62% of the cue effect, and remains a sensitivity analysis because the bounds were chosen post hoc.

E.6 SIMULTANEOUS LAYER-WISE BANDS

Pointwise intervals do not control the probability of a false exclusion when all layers are inspected. We therefore compute max-t simultaneous bands (Westfall & Young, 1993) by resampling paired items, preserving each item's full trajectory, and taking the 95th percentile of the largest studentised deviation across layers in each of 10,000 bootstrap resamples. The critical values are 3.04 for the cue-probe and 3.05 for the trace-conditioned readout, compared with 1.96 for a pointwise normal interval. A joint band over both readouts uses the 1,243 common items and has critical value 3.26.

Table 38: Llama position-resolved trajectory with each arm conditioned on its own trace ($n = 1{,}182$ after the frozen endpoint-agreement exclusion). Interior points use reconstructed prefixes followed by a canonical answer cue; endpoints use the original strings.

Trace fraction	Signed gap	Mean $\lvert G_i \rvert$	Sign agreement with cue
0.00	−0.357	1.074	1.000
0.25	−0.108	1.211	0.626
0.50	−0.288	1.541	0.635
0.75	−0.246	1.642	0.641
1.00	−0.089	2.081	0.633

Table 39: Llama correlational layer-band decomposition at the cue-probe and own-trace readouts. Intervals are paired-bootstrap 95% intervals.

Model	Band signed, cue → trace	Band $\lvert\cdot\rvert$, cue → trace	Band abs. change	Remainder abs. change	Band share, cue → trace
Llama (L24/25/27/30)	−0.350 → −0.241	0.853 → 1.092	+0.239 [0.157, 0.325]	+0.606 [0.532, 0.682]	0.797 → 0.520

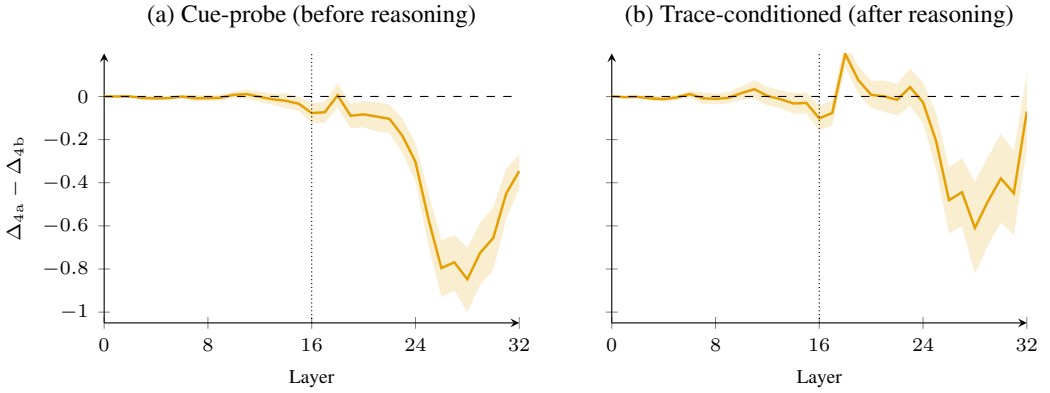

Figure 9: Layer-wise ARM 4A−ARM 4B margin difference in Llama-3.1-8B ($n = 1{,}251$ cue-probe; $n = 1{,}243$ trace-conditioned). Shading shows pointwise paired-bootstrap 95% intervals for the descriptive trajectories. The dotted lines mark L16, the first layer at which the max-t simultaneous band lies below zero for either readout. The cue-probe separation then persists across L23–32, whereas the trace-conditioned separation is narrower and shallower.

For comparison, scanning the recomputed pointwise intervals alone first yields intervals below zero at L4 for the cue probe and L16 for the trace-conditioned readout. The isolated early cue-probe exclusion does not survive family-wise correction. We therefore use pointwise intervals only to display trajectory uncertainty and base every layer-onset statement on the simultaneous bands.

For both readouts, the first negative simultaneous exclusion occurs at L16. The cue-probe then has a sustained negative block from L23 through the output layer, with its deepest separation at L28. The trace-conditioned readout has a narrower upper-layer block at L26–31. Its isolated positive exclusion at L18 lies between negative regions and is not interpreted as part of the plausible-disadvantage trajectory; identifying which component produces this post-hoc sign reversal requires a separate component-level diagnostic. Table 40 reports the cue-probe upper-layer bands.

Table 40: Llama-3.1-8B cue-probe max-t simultaneous 95% bands over the 1,251-item trajectory, shown for the sustained upper-layer region. All bounds are for the ARM 4A−ARM 4B margin difference and are negative.

Layer	24	25	26	27	28	29	30	31
Lower	-0.443	-0.765	-1.004	-0.970	-1.080	-0.948	-0.882	-0.629
Upper	-0.162	-0.373	-0.588	-0.569	-0.615	-0.500	-0.430	-0.270

F Construction and Leakage Robustness Checks

F.1 Plausible-slot dose

Let $k_i \in \{1, 2, 3\}$ be the number of plausible exemplar slots in prompt i. The frozen manifest contains 468, 525, and 258 prompts at $k = 1, 2, 3$, respectively. We fit the item-level model

$$G_i^{(\text{cue})} = \beta_0 + \beta_k k_i + \beta_n \mathbf{1}[\text{number-valued answer}] + \varepsilon_i.$$

Table 41: Llama plausible-slot dose regression for the cue-probe gap on the fixed 1,251-item mechanistic set. Here β_k is the raw-unit change associated with one additional plausible slot.

Model	β_k	Bootstrap 95% CI	HC3 SE	β_n
Llama-3.1-8B	-0.334	$[-0.447, -0.223]$	0.057	$+0.127$

The negative dose coefficient is consistent with greater suppression as more slots receive source-linked competitors. This remains a manipulation check rather than a randomised dose effect because plausible-corruption availability is determined by the retrieved exemplars.

For completeness, the fitted Llama intercept is $+0.263$. The intercept and number-answer coefficients are adjustment terms rather than confirmatory endpoints.

F.2 Answer-token length and lexical accessibility

Apart from the documented 27-item truncation exception in Appendix A, the two corrupted prompts are byte-identical outside their answer strings. We tokenise each answer line in local context using the Llama tokenizer and sum the three slot-wise length differences. Table 42 reports the equal-length comparison.

Table 42: Llama answer-token-length analysis. The cue-probe gap remains negative when plausible and unrelated answer strings have identical total token length ($n = 101$).

Model	Equal-length n	$G^{(\text{cue})}$ (95% CI)	Standardised OLS β_k, baseline → adjusted
Llama-3.1-8B	101	-0.270 $[-0.528, -0.003]$	$-0.249 \rightarrow +0.157$

The plausible and unrelated strings have exactly equal total length for 101 prompts, differ by at most one token for 315, and have a mean signed difference of -1.11 tokens. The maximum absolute difference is 22 tokens. On the exactly equal-length subset, the corresponding reasoning interaction is $D^{\text{own}} = +0.230$ (95% CI $[-0.266, +0.754]$). Its interval spans zero, so this imprecise estimate is not used as evidence. The primary length result is the cue-probe restriction.

The permuted control supplies a stronger answer-string length comparison within its 688-item Llama population. By construction, ARM 4A and ARM 4C contain the same multiset of corrupted answers in every prompt, so their total answer-line token inventory is identical apart from any in-context boundary effect; the rendered-prompt audit additionally finds identical token multisets and lengths. The resulting $G^{(4a-4c)} = -0.585$ (95% CI $[-0.677, -0.495]$) therefore cannot be attributed to differences in the corrupted strings' aggregate token lengths. It remains coupled to the original question–answer pairing and local position.

For this lexical diagnostic, every non-constant covariate is standardised before fitting. The resulting baseline coefficient is $\beta_k = -0.249$ (95% CI $[-0.332, -0.165]$), rather than the raw per-slot coefficient -0.334 in Table 41; the two numbers use the same outcome and population but different units. After adding signed and absolute length differences, answer frequency, and question-occurrence differences, the coefficient changes sign and its interval includes zero ($+0.157$, 95% CI $[-0.156, +0.297]$). The standardised question-occurrence coefficient is -0.407, approximately four times the magnitude of any other adjusted coefficient. Own-competitor construction and question occurrence are therefore coincident by design, not merely correlated. The adjusted dose coefficient cannot isolate semantic plausibility from lexical availability. The added covariates barely change in-sample fit ($R^2 = 0.031$ to 0.034), while cross-validated R^2 changes from 0.030 to 0.026.

The ARM 4C intervention supplies the direct control that this collinear regression cannot. It preserves the prompt-wide answer-string multiset while breaking recipient-specific candidate binding, with local position remaining coupled, and yields $G^{(4a-4c)} = -0.585$ (95% CI $[-0.677, -0.495]$) on the 688 eligible Llama items. The plausible effect therefore survives control for prompt-wide occurrence. This does not isolate semantic plausibility: same-question position, distance to the recipient answer line, and candidate status still move together (Appendix A.6).

F.3 TARGET OPTION ORDER AND THE UNRELATED-ARM CONTROL

We tested whether the corrupted exemplars transfer a positional schema—that the answer is the first- or second-mentioned option—to the target question. The hypothesis is not supported. We binarise mean demonstrated-rank alignment as low (≤ 0.34) or high (≥ 0.66); this excludes 242 items at exactly 0.50 and leaves 1,009 for the 2×2 diagnostic. Within each target-foil-rank stratum, higher alignment does not produce a more negative gap, and the apparent marginal association is confounded by target foil rank ($r = -0.208$).

Table 43: Llama option-order diagnostics. Panel (a) does not support positional-schema transfer. Panel (b) shows no resolved order dependence for the permuted-control $G^{(4a-4c)}$ contrast, while both contrasts involving ARM 4B depend on order.

(a) Cue-probe gap by target foil rank and schema alignment

	Alignment low	**Alignment high**
Target foil first-mentioned	-0.404 ($n = 214$)	-0.438 ($n = 397$)
Target foil second-mentioned	-0.199 ($n = 237$)	-0.128 ($n = 161$)

(b) Contrasts on the 688-item permuted-control population

Contrast	**Foil first** ($n = 421$)	**Foil second** ($n = 267$)	**First − second**
$G^{(4a-4c)}$	-0.616 $[-0.730, -0.500]$	-0.536 $[-0.685, -0.391]$	-0.080 $[-0.263, +0.106]$
$G^{(4a-4b)}$	-0.655 $[-0.798, -0.512]$	-0.274 $[-0.455, -0.097]$	-0.380 $[-0.601, -0.148]$
$G^{(4c-4b)}$	-0.039 $[-0.137, +0.059]$	$+0.262$ $[+0.118, +0.402]$	-0.301 $[-0.470, -0.129]$

On the same 688 items, clean, ARM 4A, ARM 4B, and ARM 4C margins change by 0.46, 0.47, 0.85, and 0.55 logits between foil-order strata. The larger swing is specific to ARM 4B. This attribution also holds after removing target-question lexical coincidences (Table 44).

Target-question overlap is rare: 22 of 1,251 items (1.76%; 0.64% of slots) contain any replacement string in the target question, including one replacement equal to the target gold and none equal to the foil. Rates are similar by foil rank (1.6% versus 2.1%). Removing these items strengthens the full-set ARM 4A-versus-ARM 4B order difference from -0.272 to -0.285 (95% CI $[-0.450, -0.123]$). The asymmetry is therefore not explained by this lexical coincidence. We report it as an unexplained property of the unrelated control rather than proposing a mechanism.

F.4 ANSWER MENTIONS IN GENERATED REASONING

We detect a mention only when the complete candidate appears as a contiguous token sequence, augmented by normalised numerical matching. This avoids treating the occurrence of one word

Table 44: Per-arm localisation of the option-order asymmetry. Panel (b) excludes items in which any replacement occurs in the target question. Only the unrelated arm has an interval excluding zero.

(a) Per-arm margins on the permuted-control population

Arm	Foil first	Foil second	Swing
Clean	4.382	3.926	0.46
ARM 4A plausible	2.722	2.251	0.47
ARM 4B unrelated	3.376	2.526	0.85
ARM 4C permuted	3.338	2.788	0.55

(b) Foil-first minus foil-second margin, uncontaminated items ($n = 1{,}229$)

Arm	Difference (95% CI)	Excludes zero
Clean	-0.092 $[-0.520, +0.344]$	no
ARM 4A plausible	$+0.069$ $[-0.252, +0.394]$	no
ARM 4B unrelated	$+0.354$ $[+0.017, +0.691]$	yes

from a multi-word entity as a mention of the entity. Across the 1,251 mechanistic items, the primary detector finds a gold mention in 794 traces (approximately 63%) and a foil mention in 880 (approximately 70%), where an item is counted if the candidate appears in at least one arm. In Llama, 220 of the 1,243 complete-case items contain neither candidate in either arm's generated pre-cue reasoning. Table 45 compares this post-treatment stratum with the full set.

Table 45: Llama content-plus-divergence contrast on the full complete-case set ($n = 1{,}243$) and the natural non-mention stratum ($n = 220$). Because membership is determined from generated traces, the subset estimates are post-treatment robustness diagnostics.

Quantity	Full-set estimate	Full sign p	Neither-mentioned estimate	Subset sign p
Content + divergence D	$+0.179$	5.9×10^{-4}	$+0.893$ $[+0.54, +1.27]$	2.7×10^{-6}
Cue-to-matched D	$+0.099$	0.11	$+0.032$ $[-0.16, +0.22]$	1.0
Total D	$+0.278$	4.9×10^{-7}	$+0.925$ $[+0.59, +1.29]$	3.5×10^{-5}
Cue-probe gap	-0.347	8.2×10^{-16}	-0.311 $[-0.48, -0.14]$	5.6×10^{-4}

The initial whole-word diagnostic over-counted mentions of multi-word answers when only one constituent word appeared. The primary detector instead requires the complete contiguous token sequence. Looser boundary-aware diagnostics classify 642 items as mentioning both candidates, 229 as foil-only, 142 as gold-only, and 238 as neither. After applying the primary detector and complete-case alignment, 220 items form the analysis subset.

The positive content-plus-divergence contrast therefore does not require literal occurrence of either candidate answer in the trace. Because mention status is post-treatment and may select systematically different items, its larger magnitude should not be interpreted as a population comparison. This intact-trace analysis addresses literal copying without perturbing the generated reasoning, but remains observational.

G Heads Selected on the Primary Split

Table 47 lists the 30 attention heads selected on the primary partition by the criterion in Section 5. Items are assigned to selection or evaluation halves by a deterministic hash of the question identifier. On the selection half, each head is scored by its direct contribution to the ARM 4A−ARM 4B margin difference. Within the 128 heads in layers 24, 25, 27, and 30, a head is retained when its bootstrap 95% confidence interval (10,000 resamples on the selection split, $n = 651$) excludes zero in the distractor-pulling direction. The five largest contributions ($\star$) define the carrier set. These screening intervals are not multiplicity-adjusted and are used for candidate selection, not as 128 confirmatory tests. Patching effects are measured only on the held-out evaluation split. The band is declared as configuration in the patching code with its derivation rule recorded. Recomputing this per-item ranking rule on the 651-item selection split alone returns layers 24, 25, 27, and 30 with runner-up

layer 31, matching the full set and 600-item evaluation split. The band is therefore recoverable without the evaluation half. The original value of k and the rationale for stopping at four layers were not recorded, so the band is reproducible but its cutoff remains a judgement call.

Selection is repeated under two additional partitions using $\mathrm{md5}(\texttt{seed} : \texttt{qid})$. The three partitions return 30, 31, and 26 heads. All four implicated layers are represented in every partition, and 17 heads occur in all three sets, including the five carriers used for patching (Table 46). Across partitions, the reproducible structure is an upper-layer band with a stable core rather than an exact 30-head circuit.

Table 46: Stability of Llama head selection across deterministic hash partitions. Variation among the lower-ranked heads motivates reporting a stable band and core rather than a fixed circuit.

Partition	Selected heads	Layers represented	Relation to core
Primary	30	24, 25, 27, 30	five carriers selected
Alternative 1	31	24, 25, 27, 30	five carriers selected
Alternative 2	26	24, 25, 27, 30	five carriers selected
Three-way intersection	17	24, 25, 27, 30	contains all five carriers

Table 47: Heads selected on the primary partition and their direct contributions to the ARM 4A−ARM 4B margin difference ($n = 651$). $\star$ marks the five carrier heads.

Layer	Head	DLA contribution	95% CI
24	24	-0.0504	$[-0.0769, -0.0248]$ $\star$
30	14	-0.0429	$[-0.0548, -0.0311]$ $\star$
30	11	-0.0279	$[-0.0435, -0.0127]$ $\star$
27	20	-0.0215	$[-0.0396, -0.0036]$ $\star$
27	4	-0.0190	$[-0.0295, -0.0093]$ $\star$
25	15	-0.0186	$[-0.0312, -0.0072]$
30	21	-0.0182	$[-0.0286, -0.0079]$
25	4	-0.0173	$[-0.0321, -0.0030]$
25	14	-0.0151	$[-0.0255, -0.0053]$
24	23	-0.0144	$[-0.0239, -0.0051]$
24	25	-0.0121	$[-0.0214, -0.0031]$
25	5	-0.0119	$[-0.0218, -0.0030]$
24	22	-0.0113	$[-0.0214, -0.0016]$
25	13	-0.0111	$[-0.0222, -0.0013]$
24	16	-0.0097	$[-0.0193, -0.0017]$
24	26	-0.0090	$[-0.0125, -0.0055]$
30	3	-0.0048	$[-0.0090, -0.0009]$
24	3	-0.0047	$[-0.0093, -0.0005]$
30	28	-0.0016	$[-0.0033, -0.0001]$
25	23	-0.0008	$[-0.0016, -0.0001]$
30	7	-0.0007	$[-0.0010, -0.0004]$
24	19	-0.0006	$[-0.0011, -0.0001]$
30	4	-0.0006	$[-0.0010, -0.0003]$
30	5	-0.0006	$[-0.0010, -0.0003]$
24	2	-0.0003	$[-0.0005, -0.0001]$
25	3	-0.0002	$[-0.0004, -0.0001]$
25	0	-0.0002	$[-0.0004, -0.0001]$
25	2	-0.0002	$[-0.0004, -0.0000]$
24	31	-0.0001	$[-0.0001, -0.0000]$
25	25	-0.0001	$[-0.0001, -0.0000]$

G.1 ATTEMPTED LOCALISATION BY TOKEN POSITION

Residual patching by prompt span requires token-index correspondence. The primary ARM 4A−ARM 4B pair fails this requirement for 1,151 of 1,248 span-valid items because its replacement strings have different token lengths. The remaining equal-length subset is both small

($n = 97$) and has an unresolved intact gap (-0.258, 95% CI $[-0.529, +0.019]$), so we do not interpret its positional effects.

The permuted-control ARM 4A−ARM 4C construction supplies the needed alignment: from the start of retrieved context through the answer cue, token sequences occupy identical positions on all 688 items. Appendix G.3 therefore uses this token-aligned contrast for the positional intervention, while retaining ARM 4A−ARM 4B for the full-population source decomposition. This separation prevents the underpowered 97-item subset from being treated as evidence about position.

G.2 False-induction signature of the carrier heads

We test the three-part false-induction signature of Halawi et al. (2024): attention enriched on demonstrated answer tokens, preferential attention to answer text rather than the "Answer:" cue, and an output–value circuit that promotes the tokens read at those positions. Attention enrichment is the share of a head's attention assigned to a span divided by that span's token share, so 1 denotes chance allocation. Copy score projects the head's output–value action onto the unembedding rows of the tokens read. Ten unselected heads matched to the carrier-layer multiset provide the reference.

Table 48: False-induction diagnostics for the five Llama carrier heads and ten carrier-layer-matched unselected heads ($n = 1{,}248$). All carrier–reference differences have paired-bootstrap intervals excluding zero.

Measure	Carriers, ARM 4A	Reference, ARM 4A	Carriers, ARM 4B	Reference, ARM 4B
Answer-token enrichment	1.506	0.610	2.067	0.740
"Answer:"-cue enrichment	0.486	0.842	0.511	0.940
Copy score	0.0798	0.0030	0.1451	0.0052

The carriers attend exemplar-answer tokens above chance and at roughly $2.5\times$ the enrichment of the reference heads. Their attention to the answer cue is below chance, while their copy score is positive and substantially larger than the reference. Four carriers individually satisfy all three criteria under plausible corruption: L30H14 (answer enrichment 2.26), L30H11 (1.95), L27H20 (1.50), and L27H4 (1.38), each with a resolved positive copy score. L24H24 is the exception: its enrichment is 0.437 despite being the largest direct-logit-attribution contributor. The carrier label therefore describes a shared causal role, not identical head behaviour.

Both enrichment and copy score are lower under plausible than unrelated corruption. Copying is approximately 45% weaker under ARM 4A, with the same direction in all five heads. This arm-level asymmetry does not establish that copying produces the margin difference. Copy score is defined along the demonstrated token's unembedding row, whereas the outcome is the target item's gold-minus-foil margin. Their item-level association is null:

Table 49: Item-level bridge from the false-induction diagnostics to the plausible–unrelated cue gap in Llama ($n = 1{,}248$).

Predictor	Spearman ρ with cue gap	p
Carrier-summed copy-score difference	$+0.035$	0.22
Answer-enrichment difference	$+0.015$	0.60
Individual carrier heads	-0.027 to $+0.045$	all ≥ 0.11

The aggregate copy-score and margin differences have the same sign, but the item-level analysis does not link them. The carriers are therefore false-induction-like, yet copying is not established as the mechanism that produces the margin contrast. The routed answer-line analysis below supplies a complementary null directly on the gold-minus-foil margin axis.

G.3 What the carrier heads read

All span-attention and within-question analyses use the cue-appended query position, matching the margin readout. At this position, the routed source decomposition closes against head-summed attribution.

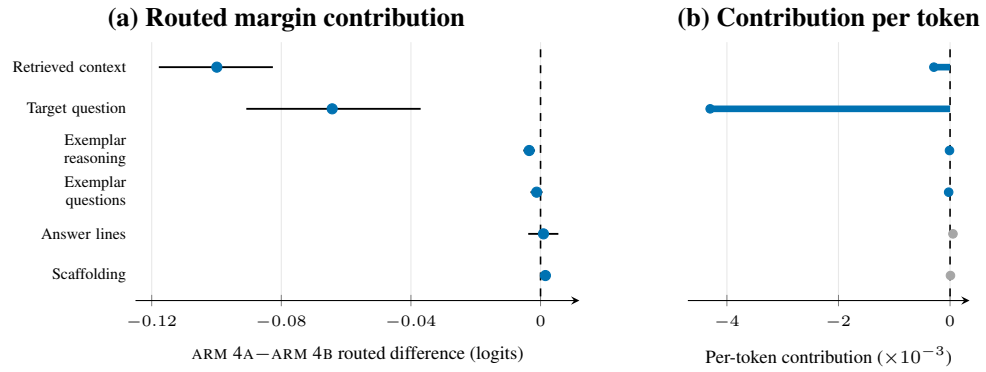

Figure 10: Where the five carriers read under the primary contrast. Panel (a) shows routed gold-minus-foil contribution by source span with paired-bootstrap 95% intervals. Panel (b) divides by mean token count: the target question supplies 39% of the carrier difference from roughly 2% of the tokens and has about 15 times the contribution density of retrieved context. Shares are of the carrier output difference, not the full margin gap.

For a token span S and carrier set H, the routed contribution at the cue-appended final query position is

$$R(S) = \sum_{h \in H} \sum_{s \in S} A_{hqs} \left\langle W_O^{(h)} v_{h,s}, d_i \right\rangle,$$

where d_i is the item-specific, final-RMSNorm-scaled gold-minus-foil unembedding direction. This quantity measures, in margin units, the part of the carriers' output routed from S. Heads are summed within each item before the arm difference is estimated. Character offsets in the rendered prompt are converted to token indices with the tokenizer's offset mapping and checked for non-empty, ordered, non-overlapping spans. Three failures are excluded, leaving 1,248 of 1,251 items.

We verify both the grouped-query-attention indexing and the decomposition itself. Llama uses consecutive key–value grouping: reconstructing each head output from attention weights and value vectors gives relative error 2×10^{-4} against `hook_z`, versus 1.30 under modular grouping. Per-position routed values close against each head's direct attribution with mean residual 1.2×10^{-4} and maximum 1.1×10^{-3}; summing spans reproduces the head-summed attribution difference at the same precision.

Table 50: Head-summed source-position decomposition under plausible versus unrelated corruption in Llama ($n = 1{,}248$). Differences are ARM 4A−ARM 4B; brackets are paired-bootstrap 95% intervals. Shares use the five-carrier total, not the full margin gap, and can be negative for opposite-signed spans. Token counts are means.

Span	Routed difference (95% CI)	Share	Tokens	Per token
Retrieved context	-0.0999 $[-0.1178, -0.0826]$	60% [51%, 72%]	345	-2.9×10^{-4}
Target question	-0.0643 $[-0.0908, -0.0370]$	39% [27%, 48%]	15	-4.3×10^{-3}
Exemplar reasoning	-0.0035 $[-0.0053, -0.0017]$	2%	345	-1.0×10^{-5}
Exemplar questions	-0.0012 $[-0.0031, +0.0006]$	1%	46	-2.5×10^{-5}
Exemplar answer lines	$+0.0009$ $[-0.0038, +0.0055]$	-0.5%	17	$+5.2 \times 10^{-5}$
Template scaffolding	$+0.0015$ $[+0.0001, +0.0029]$	-1%	198	$+7.4 \times 10^{-6}$
Five-carrier total	-0.166 $[-0.201, -0.133]$	100%	—	—

The carrier total is -0.166 (95% CI $[-0.201, -0.133]$), consistent with the attribution pipeline's -0.162 on the selection split. All five individual totals are negative with intervals excluding zero. Ten carrier-layer-matched unselected heads contribute $+0.009$ ($[+0.001, +0.017]$) in total and remain near zero on every span.

Retrieved context and the target question account for 60% and 39% of the carriers' differential contribution, respectively. The three exemplar spans account for 2% together, and the corrupted

answer lines contribute $+0.0009$—an opposite-signed share of -0.5%. The target question supplies 39% of the carrier difference from roughly 2% of the tokens; its ratio-of-means contribution density is about 15 times that of the retrieved context. Template scaffolding contributes an opposite-signed -1% share, arguing against a formatting account.

Aligned replication rules out absolute position. In the 688-item ARM 4A−ARM 4C population, token sequences from the start of retrieved context through the answer cue are position-identical. The same localisation appears: retrieved context carries 55% (95% CI $[45, 69]$), the target question 43% ($[29, 55]$), and exemplar spans 2%. Each share lies within its ARM 4A−ARM 4B counterpart's interval, so the target-side localisation is not explained by the absolute-position shifts in the primary pair.

Table 51: Source localisation across the primary and token-aligned contrasts. Brackets are paired-bootstrap 95% intervals. Shares use the carrier total for the corresponding contrast, not the full margin gap.

Source span	**ARM 4A−ARM 4B share** ($n = 1{,}248$)	**ARM 4A−ARM 4C share** ($n = 688$)
Retrieved context	60% $[51, 72]$	55% $[45, 69]$
Target question	39% $[27, 48]$	43% $[29, 55]$
All exemplar spans	2%	2%

Attention weights, rather than values, explain the routed difference. A routed contribution multiplies an attention pattern by a value write and projects through the readout normalisation. Crossing one arm's pattern with the other's values on the token-aligned contrast separates these factors (Table 52). Target-side value writes are nearly identical across arms: mean cosine is 0.9978 on the target question and 0.9976 on retrieved context, compared with 0.35–0.64 on exemplar spans whose text differs. The final-position RMS norm is 4.7% larger under plausible corruption, matching the independent 4.9% asymmetry in the geometry analysis.

Table 52: Factor-swap decomposition of five-carrier routing for ARM 4A−ARM 4C ($n = 688$). Shares concern the carrier output difference and are correlational; rounding gives 101%.

Factor	**Contribution to -0.196 (95% CI)**	**Share**
Attention weights	-0.1678 $[-0.2115, -0.1248]$	86%
Readout normalisation	-0.0168	9%
Value content	-0.0095 $[-0.0195, +0.0008]$	5%
Interaction	-0.0017	1%

The -0.196 denominator in Table 52 is the fully closed factorial five-carrier attribution total: the four rows sum to -0.1958 before rounding. Relative to the intact token-aligned gap of -0.585, it gives a 33.5% direct-attribution share. This fully closed attribution quantity is used below; it is not an intervention effect.[2] Table 51 reports source shares within each corresponding carrier total rather than claiming that its raw total equals the factorial denominator.

Bidirectional factor patching isolates the query-driven attention pattern. We cross attention patterns with value vectors at `hook_z`, then recompute attention patterns from crossed queries and keys at `hook_pattern`. "Recover" patches an ARM 4A recipient towards ARM 4C; "degrade" patches ARM 4C towards ARM 4A. The directional effect is $(\text{recover} - \text{degrade})/2$, and the offset is their mean. All swaps cover the complete key sequence, including the altered exemplar answer lines.

The full-head, pattern-only, full-pattern, and query-only means all move 13.8–14.3% of the intact gap. At `hook_z`, the mean residual $\text{full} - \text{pattern} - \text{value}$ is -0.0000 (95% CI $[-0.0013, +0.0011]$); at `hook_pattern`, the corresponding query–key residual is $+0.0009$

[2]The independently crossed-routing total is -0.179 (30.6%), whereas the head-summed source decomposition gives -0.166 on the primary ARM 4A−ARM 4B pair. They differ because the former uses a different decomposition and the latter a different contrast; neither is substituted for the fully closed factorial total.

Table 53: Factor interventions on the five frozen carriers for aligned ARM 4A−ARM 4C prompts ($n = 688$). The intact gap is -0.585 (95% CI $[-0.676, -0.495]$). Directional effects are $(\text{recover} - \text{degrade})/2$; shares divide their means by the absolute mean intact gap, with numerator and denominator jointly resampled over 10,000 draws (seed 0).

Intervention	Hook	Directional	Share of intact gap
Full head output	`hook_z`	$+0.0808$	13.8% $[10.9, 17.1]$
Pattern only	`hook_z`	$+0.0837$	14.3% $[11.4, 17.6]$
Values only	`hook_z`	-0.0028	-0.5% $[-0.8, -0.2]$
Full pattern	`hook_pattern`	$+0.0834$	14.3% $[11.3, 17.6]$
Query only	`hook_pattern`	$+0.0807$	13.8% $[10.9, 17.1]$
Keys only	`hook_pattern`	$+0.0018$	0.3% $[-0.3, +0.9]$
Control heads	`hook_z`	-0.0031	-0.5% $[-1.1, +0.0]$
Control heads	`hook_pattern`	$+0.0031$	0.5% $[+0.2, +0.8]$

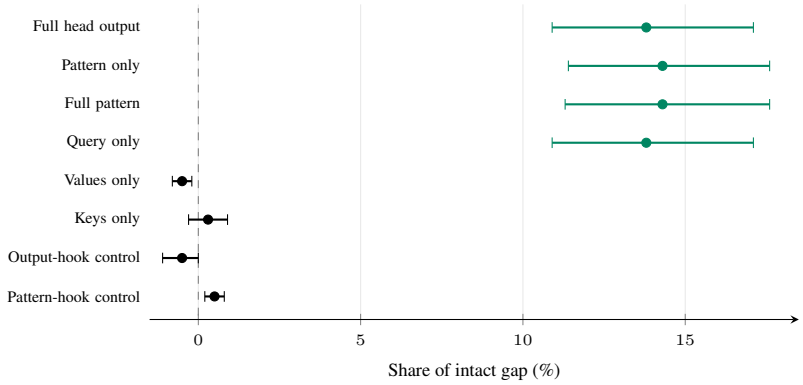

Figure 11: Reader-side factorisation on the aligned contrast in Llama. Interventions on the five carriers, expressed as shares of the intact gap with joint-bootstrap 95% intervals. Full-head, pattern, and query interventions agree in the mean, whereas key and value effects do not exceed their hook-matched control floors. This factorisation was not repeated for inconsistency.

$([-0.0004, +0.0021])$. Per-item residuals have rms 6.3% of the effect scale for both factorisations, so additivity holds for the directional means rather than item by item.

The key channel does not exceed the `hook_pattern` control floor, and the value channel is no larger in magnitude than the `hook_z` control. Their apparent raw effects lie mainly in the offset: offset-to-directional ratios are 4.4 for keys and 2.8 for values, compared with 0.10 for the full output and 0.002 for the pattern. We therefore do not interpret either null channel as mediation. The carriers were selected on ARM 4A−ARM 4B but transfer to the unseen permuted-control contrast.

G.4 Scope across ranked and upper-layer attention

Table 54 reports the bidirectional scope interventions on the five carriers, the thirty attribution-selected heads in the four-layer band, all attention heads in layers 24–31, and the residual-stream ceiling. Every share uses the same intact ARM 4A−ARM 4C gap as its denominator.

The five carriers constitute 2% of the upper-layer heads but carry about one fifth of their combined intervention effect. The thirty-head set constitutes 12% of those heads and carries about two fifths. The selected heads therefore carry a disproportionate share of an effect distributed across upper-layer attention. The remaining gap between the all-head intervention and the residual ceiling belongs to components other than upper-layer attention output at this position; it includes MLP contributions and head–MLP interactions and should not be assigned to MLPs alone.

The pattern–value result also survives expansion from five to thirty ranked heads (Table 55). At both sizes, pattern substitution reproduces the full swap in the mean, while value substitution is small and

Table 54: Causal scope on the token-aligned Llama population ($n = 688$). Each row is a bidirectional cue-position intervention expressed as a share of the intact gap (-0.585), with numerator and denominator jointly resampled. The residual row is a ceiling rather than a component set.

Intervention	Components	Share of intact gap (95% CI)
Five carriers	5 heads	13.8% [10.9, 17.1]
Selected four-layer-band heads	30 heads	27.8% [24.0, 31.9]
All attention, layers 24–31	256 heads	69.5% [64.0, 75.6]
Residual entering layer 24 (ceiling)	complete position state	99.8% [97.8, 101.9]

points in the opposite direction. The thirty-head swap independently matches the scope intervention in Table 54 to within 0.25%.

Table 55: Pattern–value factorisation at two intervention sizes. Full-swap entries are shares of the intact token-aligned gap; pattern, value, and additivity entries use the corresponding full swap as their scale. Additivity concerns directional means, not item-level effects. The thirty-head result is independently implemented and matches Table 54 within 0.25%.

Quantity	Five heads	Thirty heads
Full swap / intact gap	13.8% [10.9, 17.1]	27.7% [23.9, 31.9]
Pattern / full swap	1.035 [1.014, 1.061]	1.034 [1.017, 1.052]
Value / full swap	-0.035 [-0.061, -0.012]	-0.031 [-0.049, -0.016]
Mean additivity residual	-0.0000 [-0.0014, $+0.0012$]	-0.0004 [-0.0018, $+0.0010$]

Thirty layer-unmatched control heads move 4.3% of the intact gap (95% CI [0.3, 8.3]), roughly one sixth of the ranked-band effect. These controls lie in layers 6, 10, and 14, within the window where the query difference forms, and are therefore not a clean null at this intervention size. We interpret the result as concentration in the ranked band, not exclusivity to it.

G.5 Query direction and magnitude

We separate the five carriers' cue-position queries into Euclidean direction and magnitude. The direction-only intervention uses the donor direction with the recipient magnitude; the magnitude-only intervention uses the recipient direction with the donor magnitude. These are controlled substitutions rather than an additive decomposition.

Table 56: Bidirectional query interventions on the aligned Llama population ($n = 688$). The full row exactly reproduces the earlier query substitution. Shares use the intact cue gap and 10,000 paired-bootstrap resamples.

Query intervention	Recover	Degrade	Share of intact gap (95% CI)
Full donor query	$+0.0728$	-0.0886	13.8% [10.9, 17.1]
Donor direction, recipient magnitude	$+0.0724$	-0.0890	13.8% [10.9, 17.1]
Recipient direction, donor magnitude	-0.0007	$+0.0000$	-0.1% [-0.3, $+0.2$]
Control heads	$+0.0033$	-0.0040	0.6% [$+0.4$, $+0.9$]

The direction-only intervention reproduces the full query effect, whereas magnitude alone is indistinguishable from zero. Query norms are already closely matched: pooled across carriers, the cross-arm norm-ratio quartiles are 0.9822, 0.9959, and 1.0091, while query-cosine quartiles are 0.9812, 0.9876, and 0.9921. Table 57 gives the head-level geometry. Each carrier applies RMSNorm before W_Q, with the same learned gain in both arms, so the final-position residual-scale difference does not propagate into query magnitude.

Target-side keys and values are both nearly unchanged, whereas byte-identical later exemplar spans diverge after attending to earlier permuted answer lines. This independently illustrates absorbed state: identical text need not imply identical activations. Answer-line patching is complete by layer 8; the cue-position sweep identifies a dominant layer-15 write, and query substitution drives the

Table 57: Cross-arm query geometry at the answer cue. Each cell gives the first quartile, median, and third quartile over the 688 aligned prompts.

Head	Query cosine	Query-norm ratio
L24H24	0.9749 / 0.9818 / 0.9879	0.9792 / 0.9955 / 1.0109
L30H14	0.9804 / 0.9860 / 0.9908	0.9967 / 1.0104 / 1.0216
L30H11	0.9831 / 0.9879 / 0.9922	0.9822 / 0.9946 / 1.0057
L27H20	0.9817 / 0.9872 / 0.9917	0.9794 / 0.9912 / 1.0031
L27H4	0.9887 / 0.9919 / 0.9946	0.9793 / 0.9905 / 1.0008
Pooled	0.9812 / 0.9876 / 0.9921	0.9822 / 0.9959 / 1.0091

Table 58: Cross-arm cosine similarity of the carriers' raw key and value vectors on aligned prompts. These value cosines are computed before W_O and are distinct from the output-space value-write cosines used in Table 52, although the target-side values agree to four decimals. RoPE preserves cosine at each fixed aligned position.

Span	Raw key cosine	Raw value cosine
Exemplar answer lines	0.7553	0.3639
Exemplar questions	0.8719	0.6375
Exemplar reasoning	0.8859	0.6457
Target question	0.9985	0.9978
Retrieved context	0.9988	0.9976
Template scaffolding	0.9732	0.9250

aligned upper-reader effect. These successive endpoint interventions do not establish a direct connection. On factorial inconsistency, the receiver-specific path test instead places the writers' immediate output mainly in the layer-16 MLP and little in upper-reader queries (Appendix D.9).

Carrier attention shifts towards the target question. At the cue-appended query position, the primary contrast shifts carrier mass away from exemplar answer lines, questions, and reasoning and towards the target question (Table 59). On the token-aligned contrast, target-question mass rises by $+0.141$ while retrieved-context mass falls by -0.173; the same target-question and exemplar-span directions replicate across contrasts. These span totals are a descriptive signature: the layer-24 direction transplant reproduces them without reproducing the margin effect (Appendix D.4).

Table 59: Carrier attention-mass changes at the cue-appended query position. The primary partition calls all non-enumerated prompt tokens "remainder"; the aligned partition isolates retrieved context and template scaffolding, so the final rows are not direct counterparts. Dashes denote unavailable matched spans, not zero.

Span	ARM 4A−ARM 4B mass change	ARM 4A−ARM 4C mass change
Exemplar answer lines	-0.0455 $[-0.0496, -0.0416]$	-0.020
Exemplar questions	-0.0062 $[-0.0079, -0.0046]$	—
Exemplar reasoning	-0.0133 $[-0.0165, -0.0101]$	-0.029
Target question	$+0.1368$ $[+0.1270, +0.1467]$	$+0.141$ $[+0.131, +0.151]$
Retrieved context	—	-0.173 $[-0.188, -0.158]$
Remainder / scaffolding	-0.0717 $[-0.0823, -0.0613]$	$+0.085$

The retrieved-context values project positively onto the gold-minus-foil direction. Plausible corruption therefore makes the routed context difference negative by reducing attention to that positive content; the sign is consistent with, rather than opposed to, the context span's 55–60% share of the negative carrier difference.

Token-aligned residual patching identifies where arm information is absorbed. Bidirectional all-layer residual patching on the token-aligned contrast recovers nearly the full gap at exemplar answer lines, but not at target-side positions (Table 60). A layer sweep localises answer-line recovery

early: layer 0 alone recovers 0.569, recovery is complete by layer 8, and layers 16 onward contribute zero. Two qualifications prevent a per-span causal reading. First, patching all layers at the answer lines nearly replaces the manipulation itself. Second, the byte-identical exemplar-question span recovers 0.167, showing that residual patching measures information absorbed at a position, not only the span's lexical content.

Table 60: Residual patching by source span for the token-aligned ARM 4A−ARM 4C pair ($n = 688$). Recovery identifies where condition information has been absorbed; it is not a decomposition of the span's content.

Span patched	Recovery share of intact gap (95% CI)
Target question	-0.000 $[-0.037, +0.034]$
Retrieved context	-0.056 $[-0.095, -0.021]$
Target side, both spans	-0.033 $[-0.081, +0.010]$
Exemplar answer lines	$+0.994$ $[+0.945, +1.043]$
Exemplar reasoning	$+0.009$ (interval includes zero; one direction only)
Exemplar questions	$+0.167$ $[+0.124, +0.214]$

The localisation and patching results are complementary. Target-side values are already nearly identical, so transplanting their residual states recovers nothing; their differential carrier output instead comes from query-driven attention reweighting, established causally within the five patched heads by Table 53. The direction transplant further shows that coarse span-level reweighting is not sufficient: it can be reproduced while the margin moves in the wrong direction. All-layer answer-line patching is close to reinstating the manipulation, and the byte-identical exemplar-question recovery shows that it tracks absorbed state rather than lexical source. Source shares use carrier output as their denominator, whereas factor-intervention shares use the intact aligned cue gap; the full aligned carrier intervention moves 13.8% of that gap.

A depth probe locates when the query difference accumulates. On the 688 aligned ARM 4A−ARM 4C items, we accumulated the final-position residual by depth and mapped each partial sum through each carrier layer's query projection. Table 61 reports the arm difference in the resulting query, divided by that query's norm and averaged over the five carriers. The relative difference is small through layer 8, rises most sharply between layers 8 and 16, peaks at layer 20, and is already present before the first carrier layer. This timing independently agrees with the answer-line patching sweep: answer-line recovery is complete by layer 8 and absent from layer 16 onward, whereas the final-position query difference accumulates primarily over the intervening window. The two analyses were not designed to validate one another.

Table 61: Diagnostic query-depth probe for the aligned Llama contrast ($n = 688$). Intervals are paired-bootstrap 95% intervals. The probe applies a late-layer query map and final-depth normalisation to partial residuals, which the forward pass never does; values locate the accumulation window and must not be interpreted as causal contributions or forward-computation effect sizes.

Depth	Relative query difference (95% CI)
0 (embedding)	0.019 $[0.018, 0.020]$
4	0.026 $[0.025, 0.027]$
8	0.043 $[0.042, 0.044]$
12	0.124 $[0.121, 0.126]$
16	0.221 $[0.216, 0.226]$
20	0.240 $[0.234, 0.245]$
24 (first carrier layer)	0.227 $[0.222, 0.232]$
28	0.178 $[0.174, 0.182]$
32	0.186 $[0.182, 0.191]$

This probe alone does not identify what writes the query difference. A partial residual is not a valid input to a layer-24 query projection, and the normalisation uses the final-depth scale rather than a scale computed for that partial state. We therefore do not use it for component attribution; the separate forward-pass sweep and held-out head interventions in Appendix D.5 supply that evidence.

We omit an accompanying attention-logit ratio because its divisor lies near a sign change and the resulting values inherit the same non-forward construction.

Together, these analyses test whether corruption changes what is copied, retained, or represented from the exemplars. Most of these accounts are disfavoured. The positive evidence instead points to attention reweighting over positions whose content is nearly unchanged, although the within-question reallocation account remains unresolved.

Table 62: Candidate accounts and their status within this design. Most are disfavoured; the within-question reallocation account remains unresolved because its decisive routed-sum estimand was not reported. Tables 51–52 support attention reweighting over nearly unchanged target-side values.

Account	Status within this design
Answer-line copying	No plausible answer exactly equals the target foil; the 18 substring-containment cases are audited separately and do not remove the contrast.
Prompt-wide answer-string occurrence	ARM 4C preserves the prompt-wide answer-string multiset.
Positional schema transfer	The predicted foil-rank interaction is unsupported.
Target-question lexical coincidence	Only 1.8% are affected; excluding them does not remove the contrast.
Arm-specific trace divergence	Growth and decorrelation persist under one common trace.
Carrier suppression after reasoning	Trace-position mediation grows rather than falls.
Within-question carrier reallocation	Unresolved: the relevant prediction concerns the sum of the gold- and foil-routed changes, whereas the reported gold-minus-foil routed difference is sign-ambiguous (Appendix G.6).

G.6 WITHIN-QUESTION CARRIER REALLOCATION

The target-question mass increase could reflect greater routing through foil-negative content even without a shift in mass from the gold entity to the foil. We therefore divide the aligned target question into gold-entity, foil-entity, and remainder spans, hold the readout scale common, and recompute mass and routed quantities at the cue-appended query position. Three items in which an option cannot be located are excluded, leaving 685 of 688.

The sign test must account for the baseline projection of each span. The foil span already projects negatively onto the gold-minus-foil direction in ARM 4C (-0.311), so greater attention to it makes its routed contribution more negative. A gold-to-foil reallocation therefore predicts a negative change in the *sum* of the gold- and foil-routed contributions. It does not determine the sign of their difference.

Table 63: Within-question analysis for the token-aligned ARM 4A−ARM 4C contrast in Llama ($n = 685$). Changes are ARM 4A−ARM 4C; brackets are paired-bootstrap 95% intervals. Projection levels use a common readout scale. Per-token projection changes and attention-weighted routed contributions are distinct estimands.

Target-question group	Mass change	Mean per-token projection change	Mean per-token projection level, ARM 4C
Gold entity	$+0.0230$ $[+0.0154, +0.0307]$	-0.0056	$+0.342$
Foil entity	$+0.0270$ $[+0.0222, +0.0319]$	$+0.0040$	-0.311
Question remainder	$+0.0905$ $[+0.0851, +0.0960]$	-0.0003 (interval includes zero)	$+0.014$

Within-question contrast	Change (95% CI)	Mechanism prediction	Decision
Gold minus foil, mass	-0.004 $[-0.012, +0.004]$	Negative	Unresolved
Gold minus foil, routed contribution	$+0.065$ $[+0.050, +0.080]$	Sign-ambiguous	Non-diagnostic

Mass rises on both answer options by similar amounts and most strongly on the question remainder. The gold-minus-foil mass change is unresolved. The upper rows report changes in mean per-token projections, whereas $+0.065$ is the difference between attention-weighted routed contributions; the two quantities therefore do not subtract directly. Because the required gold-plus-foil routed sum is not reported, this analysis neither confirms nor rules out within-question reallocation.

H Uncertainty in Operational Patching Fractions

The mediated fraction is a ratio of the patching effect M to the intact held-out difference T. We preserve the paired item structure and recompute both quantities jointly within every one of 10,000 bootstrap resamples. Fractions are left unclipped. The operational fraction is defined only when the intact point estimate is positive. The analysis also checks for non-positive or near-zero bootstrap denominators and would suppress percentile ratio intervals in favour of the joint (M, T) distribution and Fieller intervals (Fieller, 1954) if denominator instability occurred. No such instability occurs in the Llama evaluation split.

Table 64: Joint-bootstrap uncertainty for the five Llama carrier heads on the 600-item held-out evaluation split. The ARM 4A-donor row reports the magnitude of the negative margin change so both operational ratios have the same positive orientation. Directional labels are used because these interventions do not support stronger functional claims.

Quantity	Point estimate	Paired-bootstrap 95% CI
Intact difference T	$+0.406$	$[+0.295, +0.520]$
ARM 4B-donor $\rightarrow$ ARM 4A-recipient effect M	$+0.072$	$[+0.049, +0.095]$
ARM 4A-donor $\rightarrow$ ARM 4B-recipient effect magnitude M	$+0.088$	$[+0.064, +0.114]$
ARM 4B-donor $\rightarrow$ ARM 4A-recipient ratio M/T	0.177	$[0.122, 0.254]$
ARM 4A-donor $\rightarrow$ ARM 4B-recipient ratio M/T	0.217	$[0.157, 0.304]$

H.1 Common-trace interchange at a trace-conditioned position

The cue-probe intervention establishes mediation of the pre-reasoning plausible–unrelated difference, not mediation of its signed-mean change. To test whether the movement of that signed mean towards zero arises by suppressing the five carriers, we repeat the same bidirectional interchange patch after appending one common reasoning trace to both corrupted prompts. Using a different trace for each arm would change both content and length and would not provide corresponding activation positions. The common-trace design restores a near-minimal pair but does not decompose the own-trace interaction D. We test reasoning generated under the clean, plausible, and unrelated arms in separate runs, so the conclusion does not depend on the clean trace specifically.

The carrier set, 600-item held-out split, answer-cue patch position, donor directions, unselected control, and exclusion list are fixed from the cue-probe mediation analysis. Each trace source covers all 600 items with non-empty reasoning. Cue and trace-conditioned conditions are run in the same process. The cue conditions contain no trace and therefore should agree across runs; all three independent processes reproduce a symmetrised mediated magnitude of 0.0799 and an intact signed ARM 4A$-$ARM 4B gap of -0.4058 to four decimals. This agreement provides a direct check that the runs remain comparable while the trace source changes.

Table 65: Llama interchange patching at cue-probe and common clean-trace answer-cue positions on the same 600 held-out items. Intervals use 10,000 paired bootstrap resamples. The primary endpoint is the absolute symmetrised mediated magnitude, $(\text{necessity} - \text{sufficiency})/2$.

Quantity	**Cue probe**	**Clean trace**	**Paired change**
Necessity, ARM 4B$\rightarrow$ARM 4A	$+0.0717$ $[+0.049, +0.095]$	$+0.2436$ $[+0.201, +0.288]$	—
Sufficiency, ARM 4A$\rightarrow$ARM 4B	-0.0880 $[-0.114, -0.064]$	-0.3086 $[-0.362, -0.259]$	—
Symmetrised magnitude	$+0.0799$ $[+0.057, +0.104]$	$+0.2761$ $[+0.231, +0.324]$	$+0.1962$ $[+0.154, +0.241]$
Signed intact gap, ARM 4A$-$ARM 4B	-0.4058 $[-0.520, -0.295]$	-0.2689 $[-0.457, -0.082]$	$+0.1368$ $[-0.038, +0.315]$
Mean absolute intact gap	1.060	1.723	—

Both intervention directions grow after the common trace. The increase is stable across trace sources: the paired changes differ by less than 0.007, their intervals overlap closely, and all three control changes remain small (Table 66). The signed intact mean also behaves similarly, reaching between -0.2580 and -0.2689 from -0.4058 at the cue probe. For the primary clean-trace condition, the paired change interval for the signed mean spans zero while the mean absolute item gap increases from 1.060 to 1.723. The trace-conditioned position therefore exhibits greater item-level dispersion and cancellation, not an unqualified reduction in the contrast.

Table 66: Trace-source invariance of Llama interchange patching. Each source trace is supplied to both corrupted prompts. Brackets give paired-bootstrap 95% intervals; all other settings and items are fixed.

Common trace source	Trace-site magnitude	Change from cue	Intact gap	Control change
Clean (ARM 1)	$+0.2761$ $[+0.231, +0.324]$	$+0.1962$ $[+0.154, +0.241]$	-0.2689	-0.019
Plausible (ARM 4A)	$+0.2699$ $[+0.224, +0.318]$	$+0.1901$ $[+0.148, +0.235]$	-0.2580	-0.020
Unrelated (ARM 4B)	$+0.2730$ $[+0.227, +0.321]$	$+0.1931$ $[+0.150, +0.239]$	-0.2675	-0.022

Twelve of the 600 items have the documented arm-dependent truncation exception. Excluding them changes no trace-source result materially. For the primary clean-trace condition, the remaining 588 items give 0.0798 ($[0.056, 0.104]$) at the cue, 0.2801 ($[0.233, 0.329]$) after the trace, and a paired increase of 0.2004 ($[0.156, 0.246]$).

As a secondary normalised summary for the primary condition, we jointly resample the mediated numerator and mean-absolute-gap denominator. No denominator-instability gate fires. The necessity fraction rises from 0.162 ($[0.143, 0.183]$) to 0.215 ($[0.191, 0.240]$), and the sufficiency fraction from 0.170 ($[0.150, 0.192]$) to 0.239 ($[0.215, 0.265]$). These increases are smaller than the absolute-effect increase because the denominator also grows. Across the common-trace conditions, the data provide no evidence that the cue-probe carriers are suppressed at the trace readout. This result neither decomposes the own-trace D nor identifies an alternative cause.

I Geometry Definitions and Diagnostics

Final-layer margins are computed directly from the model's output logits. Intermediate residual and component projections apply the cached final-RMSNorm transformation before projection through the unembedding (`apply_ln=True`); accordingly, r denotes the resulting normalised representation in the equations below. If h is a stored raw final-layer residual of dimension d_{model}, the recoverable RMS scale is $s = \|h\|/\sqrt{d_{\text{model}}}$. We therefore analyse

$$\widetilde{e} = \frac{h(\text{ARM 4A})}{s_{4a}} - \frac{h(\text{ARM 4B})}{s_{4b}}, \qquad \widetilde{c} = \frac{w(\text{ARM 4A})}{s_{4a}} - \frac{w(\text{ARM 4B})}{s_{4b}},$$

where w is the summed carrier output. Folding RMSNorm into W_U incorporates its learned gain but cannot incorporate the input-dependent division by s; each carrier output is therefore divided by the scale of the arm to which it contributes.

This distinction is empirically material. In Llama, the mean s_{4a}/s_{4b} ratio is 1.049 (SD 0.051) and differs from unity by more than 1% for 1,114 of 1,251 items. A raw residual difference therefore contains an arm-level magnitude component absent from the representation seen by the answer readout. We make post-RMSNorm geometry primary and report raw-space values only as a coordinate diagnostic.

For each item, let $e_i = r_i(\text{ARM 4A}) - r_i(\text{ARM 4B})$ and $d_i = W_U[g_i] - W_U[f_i]$, with r_i in post-RMSNorm space. We report the signed cosine

$$c_i = \frac{e_i^\top d_i}{\|e_i\|\|d_i\|},$$

the absolute projected-length fraction $|c_i|$, the squared fraction c_i^2, and the signed scalar projection $e_i^\top d_i/\|d_i\|$. These quantities are averaged over items after the indicated sign or absolute-value operation. Consequently, the mean signed cosine can be near zero even when the mean absolute projection is non-zero, because foil- and gold-directed projections cancel.

I.1 Llama diagnostics

For each of 512 draws and each item, the stricter null replaces the item's foil with another item's foil token of the same answer type, excluding the item's own gold and foil. Each draw produces a group mean, and empirical p-values use a plus-one correction (Phipson & Smyth, 2010). This null asks whether the observed effect is unusually aligned relative to directions formed by plausible alternative-answer tokens, rather than arbitrary vocabulary items.

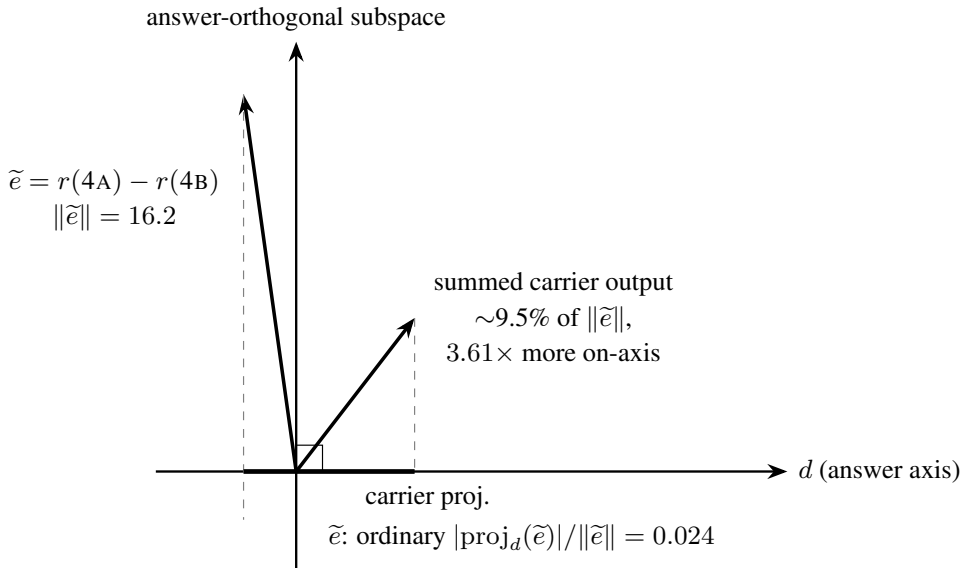

Figure 12: Llama cue-probe geometry in post-RMSNorm space. The plausible-versus-unrelated displacement is nearly orthogonal to the answer direction in absolute terms. The summed output of the five carrier heads is smaller, but its ordinary projection-length fraction is $3.61\times$ larger. The held-out primary enrichment estimate is $3.81\times$ (Appendix I). The diagram is schematic.

Table 67: Llama cue-probe geometry in post-RMSNorm space under the type-matched-foil empirical null. The enrichment divides the carriers' mean absolute cosine by that of the full displacement and uses a paired log-scale bootstrap interval.

Quantity	Full residual effect	Five carriers	Empirical test
Mean norm	16.23	0.0948 of full norm	—
Signed cosine with d	-0.00666	—	descriptive
Mean $\lvert\cos(\cdot, d)\rvert$	0.02357	0.08512	both upper-tail $p = 0.00195$
Cosine with full effect	—	0.1485	—
On-axis enrichment	—	$3.611\times$ $[3.366, 3.873]$	bootstrap 95% CI

The full displacement has small absolute answer-axis alignment: its mean absolute cosine is 0.02357. It nevertheless lies above, rather than below, the type-matched null. The carriers are also above the null and are $3.611\times$ more concentrated on the answer axis than the full displacement on the full set. Recomputing the same statistic on the frozen 600-item evaluation half gives $3.809\times$ (95% CI $[3.424, 4.211]$), compared with $3.428\times$ ($[3.120, 3.764]$) on the 651-item selection half; the bulk and carrier null tests give upper-tail $p = 0.00195$ on both halves. The held-out estimate is therefore primary. The geometric link to the patching experiment is relative concentration, not unusually low bulk alignment or strong absolute alignment of the carriers.

I.2 Raw-space coordinate diagnostic

Table 68 reports Llama raw-residual values as a coordinate-system sensitivity analysis. They are not the primary readout-space geometry because $h(\text{ARM 4A}) - h(\text{ARM 4B})$ includes the systematic difference between the two arms' RMS scales. In raw space, the full displacement lies below the type-matched null, but this reverses after the required per-arm normalisation and is therefore not interpreted. The raw full-effect signed cosine and projection are -0.00436 and -0.0913; the carrier values are -0.0278 and -0.0501.

All primary geometry measurements concern linear decodability in post-RMSNorm space at the cue-probe answer position. They do not establish a named semantic feature, causal irrelevance of the orthogonal subspace, or geometry at later reasoning positions. The analysis uses unwhitened

Table 68: Raw-residual coordinate diagnostic. These values precede the input-dependent RMS division and are not used for the primary geometric conclusions. Parenthetical tests use the type-matched-foil null.

Model	Full norm	Full mean $\lvert\cos\rvert$	Carrier mean $\lvert\cos\rvert$	Carrier/full norm	Enrichment
Llama-3.1-8B	17.003	0.02248 (lower $p = 0.00195$)	0.08621 (upper $p = 0.00195$)	0.0925	3.835 [3.573, 4.117]

representations and empirical pools matched by answer category, without additional matching on leading-space token class.

I.3 Cross-model robustness summary

Table 69: Scope of the Qwen extension. Only results reproduced in the recorded environment are included.

Analysis	Cross-model status	Supported interpretation
Consistency factorial	Replicates	Same sign, 0.297 of Llama's absolute magnitude and 0.15 of it relative to each model's clean margin; smaller main effects remain unresolved.
Residual-stream timing	Not power-matched	Qwen's residual effect first resolves at layer 18/28 versus Llama's 15/32, but Qwen has a smaller intact gap and already shows resolved layer-17 attention. This does not establish later mechanistic onset.
Arrival sublayers	Structurally different or unresolved	Qwen layer-18 attention and MLP shares both resolve; the arrival step is not cleanly attention-only.
Writer-head concentration	Open	Llama head indices do not transfer, but head identity is not comparable across models and no Qwen-specific ranking has been run.
Post-reasoning and generation	Llama only	No current-environment Qwen claim is made for reasoning dynamics, mediation, geometry, or output recovery.

J Free-Generation and Intervention Protocols

J.1 Five-arm accuracy controls

Table 70: Llama-3.1-8B-Instruct exact-match accuracy across the five arms on the 2WikiMultiHopQA behavioural set ($n = 12{,}576$). All differences from ARM 1 are significant by paired permutation and McNemar tests. Removing the reasoning (ARM 2) is the most damaging single manipulation.

Arm	Manipulation	EM	ΔEM vs ARM 1
ARM 1	clean baseline	0.552	—
ARM 2	reasoning removed	0.509	−0.043
ARM 3	answer lines blanked	0.537	−0.014
ARM 4A	plausible incorrect answers	0.516	−0.035
ARM 4B	unrelated incorrect answers	0.510	−0.042

Table 71: Llama-3.1-8B-Instruct exact-match accuracy across the five arms on HotpotQA ($n = 7{,}405$), the cross-dataset counterpart of Table 70. All differences from ARM 1 are significant by paired permutation and McNemar tests except ARM 3 ($p = 0.66$). Blanking the exemplar answer is least harmful and unrelated corruption most harmful; unlike on 2WikiMultiHopQA, unrelated corruption is slightly more damaging than reasoning removal.

Arm	Manipulation	EM	ΔEM vs ARM 1
ARM 1	clean baseline	0.510	—
ARM 2	reasoning removed	0.482	−0.028
ARM 3	answer lines blanked	0.511	+0.002 (n.s.)
ARM 4A	plausible incorrect answers	0.484	−0.026
ARM 4B	unrelated incorrect answers	0.480	−0.030

J.2 FACTORIAL FREE GENERATION

We greedily generate under all four factorial conditions with the same prompt builder, 384-token limit, and exact-match parser as the main pipeline. The four conditions for each item are decoded in one batch. Every extracted answer is classified as gold, foil, or other; no item is skipped.

Table 72: Free generation under the Llama reasoning-by-answer-line factorial ($n = 1{,}221$). Intervals are paired-bootstrap 95% intervals.

Condition	Exact match (95% CI)	Gold	Foil	Other
gg: clean	0.998 [0.994, 1.000]	1,218	1	2
gf: wrong stated answer	0.958 [0.947, 0.969]	1,170	11	40
ff: coherent wrong exemplar	0.969 [0.959, 0.979]	1,183	8	30
fg: inconsistent, correct answer	0.968 [0.957, 0.978]	1,182	16	23

Table 73: Factorial effects on exact-match accuracy, with cue-margin effects shown for scale comparison. Accuracy differences are proportions; parenthetical values give percentage points.

Contrast	Exact-match difference (95% CI)	Cue-margin difference
Answer line	−0.019 [−0.027, −0.011] (−1.9)	−0.240
Reasoning	−0.009 [−0.017, −0.003] (−0.9)	−0.110
Inconsistency	−0.020 [−0.028, −0.013] (−2.0)	−0.946
gf − gg	−0.039 [−0.051, −0.028] (−3.9)	−1.187
ff − gg	−0.029 [−0.039, −0.019] (−2.9)	−0.351
fg − gg	−0.030 [−0.040, −0.020] (−3.0)	−1.056

All three corrupted conditions lose 2.9–3.9 accuracy points relative to clean. Exact McNemar comparisons against gg give 51 versus 3 discordant outcomes for gf ($p = 2.9 \times 10^{-12}$), 38 versus 3 for ff ($p = 1.1 \times 10^{-8}$), and 39 versus 3 for fg ($p = 5.6 \times 10^{-9}$). The inconsistency and answer-line accuracy effects are both resolved and nearly equal (-2.0 $[-2.8, -1.3]$ and -1.9 $[-2.7, -1.1]$ points), although inconsistency is 3.94 times larger on the margin. Extrapolating the answer-line accuracy cost by that ratio predicts about -7.5 points, outside the observed inconsistency interval. Likewise, ff and fg cost similar accuracy despite a threefold margin difference. Of 128 errors across the corrupted conditions, 93 are non-option answers. The channel decomposition below is post hoc.

Table 74: Exact McNemar tests against the clean gg cell by error channel. Counts are corrupted-cell-only versus clean-only discordant pairs.

Contrast	Foil discordance	p	Other discordance	p
gf versus gg	11 / 1	0.00635	40 / 2	4.11×10^{-10}
ff versus gg	8 / 1	0.0391	30 / 2	2.46×10^{-7}
fg versus gg	16 / 1	0.000275	23 / 2	1.94×10^{-5}

Off-axis errors outnumber foil errors by roughly three to one in gf and ff. An exact-source audit finds no erroneous answer copied from an exemplar answer line or exemplar option in any factorial cell. Of the off-axis answers, 36 in gf, 23 in ff, and 14 in fg are spans of the target item's retrieved context; the remaining 4, 7, and 9 are unmatched. The model therefore fails mainly by selecting another contextual entity, not by copying the corrupted exemplar answer.

The pre-specified tracking rule pools the two inconsistent conditions among clean-correct items and asks whether the lower confidence bound for the AUC of margin drop predicting failure exceeds 0.5. There are 90 failures among 1,218 clean-correct items, so the pre-specified minimum of 20 failures is met.

Table 75: Association between factorial margin drops and generation failure, pooled and by error channel. The pre-specified pooled rule does not detect tracking at either readout.

Margin readout	Error channel	Failures	AUC (95% CI)
Cue probe	Pooled	90	0.500 $[0.428, 0.572]$
Cue probe	Foil only	27	0.488 $[0.361, 0.604]$
Cue probe	Other only	63	0.505 $[0.419, 0.591]$
After common trace	Pooled	90	0.527 $[0.445, 0.607]$
After common trace	Foil only	27	0.655 $[0.499, 0.802]$
After common trace	Other only	63	0.472 $[0.383, 0.557]$

The cue intervals are centred near chance both pooled and by channel. The foil-only after-trace point estimate is above chance, but its interval touches 0.5 and it rests on 27 failures, so it is an unresolved hypothesis rather than evidence of tracking. Using the bootstrap standard error implied by the pooled cue interval, the design had 80% power to detect AUC ≥ 0.60; weaker associations remain possible. No equivalence bound was fixed in advance. The clean cell is also near ceiling because the population was selected to be baseline-correct. As an anchor, its generated correctness and extracted answer reproduce the archived clean run on 99.8% of items.

J.3 Exploratory component tracking and writer-source uniformity

These follow-ups were designed after the main results were known. Their rules were fixed in script docstrings and committed before either script was first run on the real saved arrays, but the analyses remain exploratory. The polarity population contains 1,166 items, of which 1,163 supply both eligible inconsistent-cell observations. For each observation, we measure the drop from its clean partner in trace-following (b), polarity-invariant evidence (e), and the cue margin. All AUCs use identical observations and one shared 10,000-draw item bootstrap, so paired differences are computed within resample. A score is called tracking only if its 95% lower bound exceeds 0.5, and it is called better than the cue margin only if the paired-difference lower bound exceeds zero.

Table 76: Exploratory failure tracking by polarity component on the 1,163 eligible items. Counts pool the two inconsistent cells; intervals are item-bootstrap 95% intervals. No row satisfies the frozen tracking rule.

Error channel	Score	Failures / successes	AUC (95% CI)
Pooled	Trace-following b	86 / 2,240	0.499 $[0.419, 0.577]$
	Evidence e	86 / 2,240	0.567 $[0.493, 0.637]$
	Cue margin	86 / 2,240	0.498 $[0.423, 0.569]$
Foil only	Trace-following b	27 / 2,240	0.610 $[0.456, 0.748]$
	Evidence e	27 / 2,240	0.625 $[0.485, 0.760]$
	Cue margin	27 / 2,240	0.486 $[0.359, 0.606]$
Other only	Trace-following b	59 / 2,240	0.448 $[0.358, 0.533]$
	Evidence e	59 / 2,240	0.540 $[0.453, 0.622]$
	Cue margin	59 / 2,240	0.503 $[0.413, 0.591]$

Table 77: Paired AUC differences against the cue margin under the same resamples. Every interval includes zero.

Error channel	$b-$cue (95% CI)	$e-$cue (95% CI)
Pooled	$+0.001$ $[-0.088, +0.091]$	$+0.069$ $[-0.007, +0.146]$
Foil only	$+0.124$ $[-0.027, +0.267]$	$+0.139$ $[-0.000, +0.294]$
Other only	-0.055 $[-0.165, +0.053]$	$+0.037$ $[-0.052, +0.126]$

The component that carries most of the inconsistency effect therefore predicts failure no better than chance and does not outperform the original margin. The cue AUC differs trivially from Table 75 because this analysis is restricted to items with a valid polarity read; the primary 0.500 estimate remains the headline result.

We separately compare the per-item share of the writers' differential output originating at the manipulated answer span between failures and successes, conditioning on a correct answer in the consistent partner cell. The frozen rule returns *uniform* because the answer-span difference interval includes zero in both factorial pairs. For fg$-$gg there are 21 failures and 562 successes, with a failure-minus-success difference of $+0.011$ $[-0.030, +0.049]$; for gf$-$ff there are 15 failures and 550 successes, with a difference of -0.065 $[-0.145, +0.003]$. These are means of per-item shares and therefore need not equal the ratio-of-means value in Table 20. Other span comparisons were exploratory and are not interpreted. Whatever determines which items fail is not resolved by how strongly these writers differentially read the corrupted answer span.

J.4 Margin, accuracy, and error composition

The five-arm tables and cue-probe margins answer different questions: exact match evaluates the complete generated answer, whereas the margin compares only the gold and foil first tokens. We align the two readouts on all 1,251 mechanistic items with auditable outcomes under both corrupted arms. Table 78 gives the paired outcome table.

Table 78: Paired Llama exact-match outcomes on the 1,251-item mechanistic set. Plausible-corruption accuracy is 93.8% and unrelated-corruption accuracy is 91.3%, a risk difference of $+2.5$ percentage points (exact McNemar $p = 0.0035$).

	Unrelated corruption (ARM 4B)	
Plausible corruption (ARM 4A)	**Correct**	**Incorrect**
Correct	1,104	69
Incorrect	38	40

Among the 107 discordant items, the plausible-minus-unrelated cue margin averages -0.124 when only ARM 4A is correct and $+0.093$ when only ARM 4B is correct. The between-group difference

is -0.218 (95% bootstrap CI $[-0.772, +0.309]$; permutation $p = 0.42$), so the cue margin does not order the winning arm. The margin measures competition along the pre-specified gold–foil direction; it does not summarise all routes to an incorrect full answer.

An independent exact-string audit classifies each generated answer by its relation to the prompt (Table 79). It tests the direct-copy account against 2,292 inserted-string candidates in each corrupted arm, or 1.83 candidates per item. No erroneous answer in either arm is an exact copy of an inserted string. One ARM 4B answer matches an inserted string, but that string is also the gold answer or foil and therefore does not belong to the other-entity surplus.

Table 79: Exact-string answer audit on the 1,251-item mechanistic set, using the sharded generation runs behind the reported accuracies. Categories are mutually exclusive under the audit's priority order. The gold counts reconcile exactly with Table 78.

Generated-answer category	ARM 1	ARM 4A	ARM 4B
Gold	1,251	1,173	1,142
Foil	0	17	16
Inserted corruption string	0	0	0
Other entity in retrieved context	0	55	79
Other entity elsewhere in prompt	0	1	1
Entity absent from prompt	0	5	13

The other-entity error count is 61 under plausible corruption and 93 under unrelated corruption. Of the 32-error surplus, 24 answers are entities from the retrieved context and eight are absent from the prompt. The unrelated arm therefore does not supply a string that the model later copies; it more often disrupts selection among contextual entities. Foil counts remain nearly identical despite the plausible arm inserting the competitor into every corrupted exemplar. This rules out exact reproduction, not more diffuse influence from the inserted strings.

A later regeneration yields 1,172 plausible-arm and 1,140 unrelated-arm gold answers, differing in correctness on five items from the reported runs. None of those items produces a foil or an inserted corruption string, so foil parity and the exact-copy null hold in both passes; the other-entity surplus is 32 in the reported runs and 33 in the regeneration.

J.5 Free generation under corruption

The model greedily decodes the full reasoning and answer (`do_sample=False`) with repetition penalty 1.0 and at most 384 new tokens. The prompt builder, decoding parameters, and answer parser are the same as in the stored clean run. All counts in this subsection are computed from the archived release snapshot used by the accompanying analysis. In a 30-item clean-path fidelity check, 29 regenerated reasoning traces match the stored traces exactly; we do not attribute the remaining discrepancy to a specific source. Reasoning change is assessed by exact string comparison; whitespace normalisation gives the same classifications.

Table 80: Llama-3.1-8B free-generation outcomes from the full corrupted-arm runs on the 1,251-item baseline-correct mechanistic set. The changed-trace column conditions on a post-treatment variable and is descriptive.

Corruption	Changed traces	Change rate	Overall correct	Correct among changed
Plausible (ARM 4A)	472/1,251	37.7%	1,173/1,251 (93.8%)	394/472 (83.5%)
Unrelated (ARM 4B)	509/1,251	40.7%	1,142/1,251 (91.3%)	400/509 (78.6%)

Under plausible corruption, the 779 unchanged traces retain the clean answer by construction because the mechanistic items are baseline-correct. This subgroup provides no independent evidence of repair. Of the 472 changed traces, 394 retain the correct answer and 78 do not, giving a conditional failure rate of 16.5%. The relevant observation is that many textually changed traces still accompany the correct answer. The 83.5% and 78.6% values are post-treatment descriptions rather than unconditional causal repair rates; the direct paired accuracy comparison is reported in Appendix J.4.

J.6 TRACE-CONDITIONED MARGIN AND GENERATED CORRECTNESS

We test whether the Llama trace-conditioned readout is associated with the model's answer under plausible-corruption free generation. The stronger analysis restricts attention to changed traces, so every included item exhibits a behavioural response to the corruption. From the 1,251 mechanistic items, 472 traces change: 394 retain a correct answer and 78 do not. All 394 successes and 73 failures have a valid trace-conditioned margin, leaving 467 complete cases. The predictor is the final-layer ARM 4A gold-minus-foil margin. We compare groups with a two-sided, tie-corrected Mann–Whitney test and use 10,000 stratified-bootstrap resamples under project seed 20260902 for the mean-difference interval, resampling the success and failure groups separately at their observed sizes because the estimand compares their conditional means.

Table 81: Association between the trace-conditioned margin and correctness under plausible-corruption free generation in Llama. The changed-trace restriction is the primary comparison; the pooled result is reported as a weaker whole-population analysis.

Quantity	**Value**
Changed-trace complete cases	467 (394 correct / 73 incorrect)
Mean margin, correct / incorrect	9.640 / 7.042
Changed-trace difference (95% CI)	$+2.599$ $[1.184, 4.047]$
Mann–Whitney U; two-sided p	17,315; 0.0056
Pooled complete cases	1,243 (1,170 correct / 73 incorrect)
Pooled mean margin, correct / incorrect	9.685 / 7.042
Pooled difference (95% CI)	$+2.643$ $[1.247, 4.086]$
Pooled two-sided p; point-biserial r	0.0024; 0.158

The pooled complete-case analysis includes 1,243 items: 1,170 correct generations and 73 failures. Its difference can partly reflect the 779 unchanged traces, which are necessarily correct and may be easier items. Restricting to changed traces removes that particular explanation while retaining the same direction. In both analyses, high-margin failures and low-margin successes occur. Generated correctness remains a post-treatment outcome, and the comparison is an association between measured quantities rather than a causal effect of the margin. The changed-trace analysis was frozen before execution but was not pre-specified for the original study.

J.7 GENERATION-TIME PATCHING

The intervention set comes from a separate regeneration of the 1,251 mechanistic items and consists of 79 cases that are correct under the clean baseline, incorrect under plausible-corruption generation, and textually changed in their reasoning. The archived failed-repair identifiers match that regeneration exactly. Relative to the full run used for the behavioural and margin analyses, 32 plausible-arm traces and three exact-match outcomes differ; the regeneration therefore contains 79 failed repairs where the full-run funnel contains 78. The recipient is regenerated under ARM 4A. At the final prompt position and every generated position, selected heads' outputs are replaced with values from a clean-context donor. At each step the donor is evaluated on the clean prompt followed by the recipient's realised generated tokens, aligning runs by generated-token index despite different prompt lengths. Recovery is generated exact match.

A no-op rerun measures numerical stability, with a pre-specified validity threshold of at most 5% flips. The realised flip rate is 3.8%, excluding three items and leaving 76. The primary comparison patches the five carriers or a count- and carrier-layer-matched non-selected set and uses a one-sided exact McNemar test at $\alpha = 0.05$. Under the pre-specified decision rule, failure to outperform the control counts as a negative primary result. Two pre-specified diagnostics patch the full primary-split selected set and an early-layer set drawn from layers 6, 10, and 14.

The early-layer diagnostic recovers 0 of 76 items, the carriers 5, the pre-specified carrier-layer-matched control 17, and the full selected set 18. The carrier-versus-control test gives $p > 0.99$, so the negative primary decision is unchanged. Each control matches the carriers' layer multiset: one head from layer 24, two from layer 27, and two from layer 30; layer 25 is absent because no carrier comes from that layer. We draw ten further five-head sets under the same rule, selected-head

Table 82: Llama-3.1-8B generation-time patching on 76 determinism-stable failed-repair items; three no-op-unstable items are excluded. Carrier patching does not outperform the pre-specified carrier-layer-matched control. Subsequent carrier-layer-matched draws characterise the variability of this control in Table 83.

Condition (generation-time patch towards clean donor)	Recovered	Rate	Role
Early-layer heads (layers 6, 10, 14)	0/76	0.000	inert control (validity)
Carrier heads (top-5)	5/76	0.066	primary
Carrier-layer-matched random (L24/L27/L30)	17/76	0.224	active-band control
Primary-split selected set (30 heads)	18/76	0.237	band diagnostic

exclusions, and seeds 1000–1009, and patch them on the same 76 items. The carrier and diagnostic conditions from the initial run are read from their archived rows rather than rerun.

Table 83: Recovery across ten additional carrier-layer-matched, five-head control draws. Six draws recover no items; the complete count sequence is $[1, 0, 6, 0, 0, 0, 0, 0, 3, 1]$.

Carrier-layer-matched summary	Recovered	Rate	Role
Median across ten draws	0/76	0.000	null distribution
Mean across ten draws	1.1/76	0.014	null distribution
Range across ten draws	0–6	—	null distribution

The carriers recover more items than nine of ten subsequent draws, whereas the pre-specified control and the 30-head diagnostic exceed all ten. With the plus-one correction, ten draws cannot yield an empirical p below $1/11 = 0.091$; we therefore treat the distribution descriptively. Across all patched configurations, every recovered item belongs to a pool of 22: the pre-specified control and 30-head diagnostic recover 17 and 18 items, overlap on 13, and have a union of 22; the six items reached by any new draw are a subset of that union. Thus 54 items are never recovered. Recovery is concentrated in a minority of items and depends on head configuration rather than being distributed uniformly across the band.

Because the initial run used a different software environment, we replicated its protocol on 12 items before interpreting the additional controls. No-op text identity against stored ARM 4A generations is 1.000, the no-op flip rate is 0/12, and the carrier and pre-specified control recoveries are reproduced exactly (4 and 3). Per-item outcomes match on all 12 items under both conditions, supporting direct comparison with the initial run.

K Cross-Dataset Construction Density

The behavioural five-arm design transfers directly to HotpotQA, but the mechanistic plausible-versus-unrelated contrast requires an exemplar-specific competing option. This is common in 2WikiMultiHopQA comparison questions and uncommon in HotpotQA's bridge-dominated distribution. Table 84 quantifies the resulting dilution.

Table 84: Density of the plausible construction across datasets. HotpotQA supports the behavioural replication but a substantially weaker mechanistic contrast.

Construction statistic	2WikiMultiHopQA	HotpotQA
Targetable chains in full database	3,113/12,576 (24.8%)	659/7,405 (8.9%)
Plausible retrieved slots among eligible items	61.1%	32.6%
Items with source-linked competitors in all three slots	258	20
Approximate paired mechanistic items	1,251	190

Under a linear slot-density heuristic, the HotpotQA ARM 4A−ARM 4B contrast would be approximately half the -0.35 cue-probe difference measured on 2WikiMultiHopQA, while roughly 190 paired items yield an estimated bootstrap half-width near 0.2. This calculation is a design diagnostic rather than an empirical transfer result: a null mechanistic result at this density would conflate failure

to transfer with insufficient contrast. Extending the mechanistic analysis requires a new plausible-alternative construction for bridge questions rather than a direct rerun of the same pipeline.

L Reproducibility and Claim-Scope Checklist

Table 85: Reproducibility checks for data construction, population definition, and the behavioural analyses. The scripts operate against the archived data snapshot described in the reproducibility statement.

Check	Purpose
Context retrieval	Verifies the KG^{2}RAG export provenance, 12,576 unique question identifiers, and non-empty consumed fields.
Exemplar retrieval	Verifies the encoder, 1,024-dimensional normalised vectors, candidate pool, cosine ranking, top-three selection, and leave-one-out exclusion.
Cross-arm retrieval identity	Confirms the same exemplars, passages, and exemplar order in ARM 4A and ARM 4B.
Reasoning-truncation audit	Reconstructs the 353 above-threshold slots, 309 differing-length cases, and 27 affected slots and items, then reproduces the restricted-sample sensitivity estimates.
Slot-strategy audit	Reproduces 2,292 source-linked and 1,458 fallback slots among 3,750 valid slots.
Plausible-occurrence audit	Confirms exact byte occurrence of all 2,292 plausible competitors in their own exemplar questions under both model tokenizers.
ARM 4C construction audit	Reproduces the compatibility graph, selected cycle cover, deterministic tie-break inputs, screen rejections, and 783/688/638 population accounting.
ARM 4C multiset equality	Confirms per-prompt equality of answer strings and tokens between ARM 4A and ARM 4C after rendering and truncation.
ARM 4C prompt invariants	Checks equal rendered length, byte-identical non-answer text, stable truncation and occurrence, no self-assignment, and no inserted target entity.
ARM 4B occurrence audit	Measures 1.55% slot-level contamination using unchanged prompt text outside corrupted answer spans.
Target-question contamination audit	Measures the 22 affected items, gold/foil coincidences, foil-rank balance, and the exclusion sensitivity of the option-order asymmetry.
Three-arm recapture concordance	Recomputes the permuted-control per-arm and gap-level gates for the pre-specified Llama decision rule.
Recorded-environment anchors	Reproduces Llama exactly and the Qwen factorial within fp16 tolerance.
Answer-line leakage audit	Reconstructs the 18 mechanistic items whose ARM 4A exemplar answer strings contain the target gold or foil.
Plausible-dose audit	Reproduces the 468/525/258 prompt counts and the Llama dose coefficient with its paired-bootstrap interval.
Layer-band provenance	Recomputes the implicated attention band on the full, selection, and evaluation sets and returns layers 24, 25, 27, and 30, with layer 31 as runner-up.
Population funnel	Reconstructs the 12,576, 3,040, 1,251, 1,243, 1,221 factorial, 584 held-out factorial, 79, and 76 item sets.
First-token validity	Confirms distinct gold and foil first tokens for every mechanistic item.
Length and mention checks	Reconstructs each tokenizer's equal-length subset and the Llama natural non-mention subset without proxy substitution.
Matched-trace audit	Verifies donor identity, type agreement, answer exclusion, realised length balance, and donor reuse against the archived target–donor mapping.
Matched-trace dependence	Recomputes item-level and donor-clustered intervals for the archived two-way contrasts, leave-one-donor-out shifts, and the permuted-donor question-similarity null.
Three-way trace decomposition	Verifies shared matched and common traces, exact item-level closure, and the separate matched-irrelevant-insertion, fixed-trace-content, and divergence steps.
Placebo baseline sensitivity	Stratifies each paired contrast by its own cue-$\lvert G\rvert$ quartiles, recomputes within-stratum growth excesses, and inspects their cue-magnitude sensitivity without selecting on only one denominator.
Exact-match normalisation	Recomputes all 12,576 clean outcomes with the archived parser and reproduces 6,937 stored correct answers exactly.
Factorial free generation	Regenerates all four factorial cells in matched batches, recomputes exact-match and answer categories, factorial accuracy contrasts, exact McNemar tests, and cue- and trace-margin failure AUCs.
Factorial error channels	Splits foil and off-axis failures, runs exact McNemar tests by channel, audits every off-axis answer against exemplar lines, options, and target context, and recomputes channel-specific AUCs.
Polarity-component tracking	Reuses the 1,166-item polarity capture, applies one shared item bootstrap to b, e, and cue-margin AUCs, and recomputes paired component-minus-margin differences on identical observations.
Writer source by outcome	Compares per-item manipulated-answer-span shares between failures and successes in both factorial pairs, conditioning on correctness in the consistent partner cell.

Table 86: Reproducibility checks for mechanistic, robustness, and generation analyses (continuation of Table 85).

Check	Purpose
Simultaneous bands	Recomputes max-t critical values jointly over the layer trajectories.
Patching-fraction uncertainty	Jointly resamples the mediated numerator and intact-gap denominator on the held-out split.
Factor-share recomputation	Expresses all `hook_z` and `hook_pattern` interventions against the aligned intact gap using 10,000 joint-bootstrap draws with seed 0.
Common-trace interchange	Repeats the trace-position patch under clean-, plausible-, and unrelated-arm trace sources and reproduces identical cue-probe magnitudes and intact gaps to four decimals across the three independent processes.
Grouped-query-attention mapping	Reconstructs each carrier output from attention weights and value vectors and distinguishes the correct consecutive key–value grouping from the incompatible modular mapping.
Cue-position mapping	Recomputes carrier span attention and within-question reallocation at the cue-appended query position and verifies closure against head-summed attribution.
Source-position closure	Verifies that per-position routed values close against direct head attribution and that source spans reproduce the head-summed carrier difference at the cue-appended query position.
Head-summed source decomposition	Recomputes routing over context, target-question, exemplar, and template spans for ARM 4A−ARM 4B and the token-aligned ARM 4A−ARM 4C population, with carrier-layer-matched references.
Consistency factorial	Rebuilds 2,150 eligible slots, verifies byte-identical gg/ARM 1 and gf/ARM 4A anchors, and recomputes condition margins, factorial contrasts, flip-complete sensitivity, and scaling fits.
Qwen consistency factorial	Reuses the frozen construction manifest on all 1,221 items and recomputes the factorial contrasts.
Qwen arrival sweep	Patches all 28 residual-stream inputs, reruns layers 15–21 independently, and obtains bit-identical residual effects in the overlapping window.
Factorial own-trace readout	Reads every factorial cell after its own generated reasoning on 1,209 complete items and compares each contrast with the matched common-trace readout.
Trace-polarity decomposition	Reconstructs the swapped traces, gate exclusions, clean-channel scales, manipulation check, and evidence-versus-trace-following contrasts on 1,166 items.
Direction transplant	Recomputes selection-split mean directions, held-out bidirectional effects, placebo and random-direction controls, and the pre-specified layer-24 portability decision.
Cue-position arrival sweep	Repeats per-item residual, attention-output, and MLP-output interchange through layers 12–16 and verifies agreement with the independently run transplant anchors.
Layer-15 head writers	Reconstructs the selection ranking and held-out top-k, complement, random-set, and additivity interventions on disjoint item splits.
Inconsistency writer transfer	Reconstructs the factorial evaluation split, bidirectional arrival and sublayer sweeps, and the no-reselection comparison of the aligned writer set with its complement and random sets.
Inconsistency reader transfer	Applies the frozen carrier, ranked-band, all-upper, random, and residual-ceiling sets without reselection, recomputes both factorial pair directions, and checks the recorded 0.113 maximum cross-environment item drift.
Writer source decomposition	Reconstructs writer `hook_z` from disjoint prompt spans, verifies maximum closure error 2.40×10^{-4}, and recomputes answer-span and template shares.
Writer-direction decomposition	Estimates each factorial-pair direction on 200 selection items, evaluates projected, orthogonal, and norm-matched random interventions on 584 held-out items, and verifies equivalence to wholesale writer patching.
Writer-direction controls	Recomputes pair-direction agreement, per-item coherence, answer-axis alignment, clean-output cosines, and centred and uncentred top-five subspace projections against in-span references.
Writer-to-receiver paths	Delivers the frozen writer difference to one query or MLP receiver at a time, verifies query reconstruction error, and retains unpatched keys and values.
Specificity controls	Applies frozen writer and carrier sets to reasoning removal, answer blanking, and answer corruption, then records the pre-specified mixed verdicts and unresolved context perturbation.
Attention–value factor swap	Crosses each aligned arm's attention patterns, value writes, and readout scale to reproduce the 86%/9%/5% carrier-output decomposition and target-side value cosines.
Bidirectional pattern–value patch	Rebuilds per-query-head outputs from crossed patterns and values, checks self-reconstruction error, reports both donor directions and offsets, and verifies directional additivity.
Bidirectional query–key patch	Recomputes cue-position patterns from crossed queries and keys, reports both directions, verifies query–key closure and raw key/value cosines, and cross-checks the full pattern effect against the independent `hook_z` implementation.
Query direction–magnitude patch	Reproduces the full query swap, substitutes direction and magnitude separately, and verifies the RMSNorm account of the null magnitude effect.
Intervention scope hierarchy	Repeats output patching on five carriers, the thirty attribution-selected heads, all heads in layers 24–31, and the residual entering layer 24.
Thirty-head factor repeat	Repeats the pattern–value intervention on the ranked band and verifies agreement with the independently constructed band swap.
Aligned residual patching	Recomputes bidirectional all-layer span patches and the answer-line layer sweep on 688 position-identical ARM 4A−ARM 4C items, retaining the absorbed-information caveat.
Query-depth probe	Recomputes the partial-residual query-norm trajectory on the same 688 aligned items, records the excluded attention-logit readout, and checks its independently estimated accumulation window against the answer-line patching sweep.
Within-question reallocation	Recomputes gold-entity, foil-entity, and remainder mass and routed projections at the cue-appended query position on 685 aligned items, with the pre-specified sign requirement.
Common-trace contrast comparison	Recomputes cue, own-trace, and identical-common-trace signed means, absolute gaps, rank correlations, implied and observed sign agreement, and shrinkage fits on one 1,243-item plausible–unrelated population.
Permuted-control common-trace recapture	Compares ARM 4A−ARM 4B and ARM 4A−ARM 4C in one process on the same 688 items, verifies exact alignment, and reproduces the contrast-wise magnitude, direction, signed-mean, and corruption-count results.
Unrelated-arm placebo	Verifies the preserved replacement multiset, same-form derangement, absence of self-assignment, 91.2% changed slots, and at least one changed slot in every prompt; jointly bootstraps the ratio-of-means comparison on the 684-item paired intersection.

Table 87: Reproducibility checks for robustness and generation analyses (continuation of Tables 85 and 86).

Check	Purpose
Position-localisation eligibility	Reproduces the 1,151 primary-pair token-index mismatches and verifies position identity from context through cue for all 688 aligned occurrence-control items.
False-induction diagnostics	Recomputes answer-token and cue enrichment, output–value copy scores, carrier-layer-matched references, per-head signatures, and the null item-level bridge to the cue gap.
Endpoint gap dynamics	Recomputes signed means, mean absolute gaps, margin-normalised gaps, shrinkage-model fits, and deadbanded sign concordance from the paired cue and own-trace margins.
Target option-order audit	Reconstructs foil rank and alignment bins, reproduces the unsupported 2×2 schema-transfer test, and localises order dependence to contrasts involving the unrelated arm.
Position-resolved trajectory	Reconstructs five trace-prefix readouts in per-arm and common-trace conditions, verifies both endpoint anchors, and reproduces the frozen exclusion sensitivity.
Layer-band readout decomposition	Decomposes cue and own-trace gaps into implicated-band and remainder contributions with exact item-level closure.
Geometry normalisation and null	Recomputes answer-axis geometry from the archived capture in post-RMSNorm space using the same analysis script and 512-draw type-matched-foil null, and reports the per-arm RMS-scale-ratio diagnostic.
Held-out geometry split	Recomputes carrier enrichment and bulk-null tests on the frozen 651/600 selection and evaluation halves.
Behavioural regeneration	Reproduces exact-match totals and changed-trace rates from stored generations.
Margin–generation association	Reproduces the 1,243 pooled complete cases, 467 changed-trace complete cases, the 73/78 margin-side failure accounting, and all reported association statistics.
Margin–accuracy reconciliation	Reconstructs the paired 1,251-item accuracy table, discordant-item margin test, and foil-versus-other-entity error counts.
Clean-path fidelity	Regenerates 29 of 30 sampled clean reasoning traces exactly.
No-op stability gate	Excludes three generation-time items and verifies the 3.8% flip rate is below 5%.
Generation-time protocol replication	Reproduces the initial carrier and carrier-layer-matched-control outcomes on 12 items, with no-op text identity 1.000 and flip rate 0.000.
Figure regeneration	Recreates all plotted coordinates from the archived analysis outputs.
Evidence-tier language	Restricts decodability and attribution claims to correlational evidence and causal language to interventions.

L.1 Analysis-environment manifest

The final model-inference environment uses PyTorch 2.11.0+cu128, Transformers 5.16.1, TransformerLens 3.9.0, and huggingface-hub 1.29.0 in fp16 with greedy decoding and the library-default attention implementation. The recorded model revisions are `0e9e39f2` for Llama-3.1-8B-Instruct and `a09a35458c702b33eeacc393d103063234e8bc28` for Qwen2.5-7B-Instruct. Unless otherwise stated (Reproducibility Statement), seed 20260902 controls the 10,000-draw item bootstrap, not inference.

The component-tracking and writer-source outcome follow-ups run only on the saved second-batch arrays and require no new model inference or GPU computation. Their scripts and decision rules were committed before first execution on the real arrays; their post-results design is retained explicitly in Appendix J.3.

Table 88 records the cross-session checks. Every reported capture is recomputed in the final environment and compared item by item: Llama agrees exactly over 80 comparisons, and Qwen within the last fp16 digit over 120. Independently rerunning the layers 15–21 Qwen window reproduces the overlapping residual effects from the full-depth sweep bit for bit.

Table 88: Recorded-environment reproducibility checks. Each reported capture is recomputed in the final environment and compared item by item.

Diagnostic	Llama	Qwen
Recorded model revision	`0e9e39f2`	`a09a3545...`
Shared comparisons	80	120
Mean absolute difference	0.0000	0.0006
Maximum absolute difference	0.0000	0.0078
Independent arrival-window residual rerun	—	maximum difference 0.000000
Interpretation	bit-identical	fp16 tolerance

Greedy margin captures are deterministic within a fixed environment. The project seed controls bootstrap resampling, not model inference, unless stated otherwise (Reproducibility Statement). The release artefact set is assembled from the preserved backup snapshot because three source com-

ponents are incomplete in the working copy: `database_v2/full_records.jsonl` and the sharded `r001_arm4a` and `r001_arm4b` generation runs behind the reported accuracies. The working-directory copy of `r001_arm4a` has lost its generation shards, while the `_gen` directories contain the later regeneration rather than the reported pass. The released backup copies are hashed with the rest of the artefact set.

Table 89: Recorded analysis and inference environment for the final Llama and Qwen results.

Component	**Recorded analysis environment**
Model identifiers	`meta-llama/Llama-3.1-8B-Instruct`; `Qwen/Qwen2.5-7B-Instruct`
Model revisions	Llama `0e9e39f2`; Qwen `a09a3545...`
Tokenizer identifiers	Same repositories and revisions as the corresponding models
PyTorch / Transformers	2.11.0+cu128 / 5.16.1
TransformerLens / huggingface-hub	3.9.0 / 1.29.0
NumPy / SciPy / statsmodels	2.1.3 / 1.16.3 / 0.15.0
CUDA / cuDNN	12.8 / 91900
GPU	NVIDIA A100-SXM4-80GB $\times$1 (79.3 GB)
Inference dtype / attention	float16 / library default (not explicitly set)
Decoding	greedy (`do_sample=False`); temperature not passed; repetition penalty 1.0; maximum 384 new tokens
Seeds (bootstrap / generation)	20260902 / not used under greedy decoding
Analysis commit	`a9891e2f33aa`