WHAT DOES A HALLUCINATION SCORE TELL US? AUDITING THE EVIDENCE FROM PREDICTION TO MECHANISTIC CLAIMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Internal representations offer a promising basis for detecting hallucinations in large language models and investigating how errors arise. Yet a score that predicts errors does not by itself establish which computation produced them or where an intervention should act. Connecting these claims requires tracing whether the observation, predictive support and intervention endpoint refer to the same target. We introduce a measurement-to-claim audit organized around three properties: risk observability, conditional quantitative readoutability and support identifiability. A separate intervention audit checks whether the output endpoint discriminates the target relation before testing selected-support effects against matched controls. Exact-prefix checks, candidate-conditioned readouts and support resampling establish the scope of predictive evidence; a controlled nonlinear calibration distinguishes known semantic-path supports from predictive proxies and response biases. Our main natural-model audit evaluates frozen supports on the same sources across four answer codebooks and both semantic mappings. Gemma retains internal predictive information across mappings, but output-endpoint discrimination and intervention specificity require separate evidence: even where both mapped endpoints discriminate the relation, a selected-support advantage over controls remains unestablished. A source-matched Qwen stress test stops earlier because none of its evaluated output endpoints reliably discriminates the relation, despite weaker internal readability. These results make the gap between prediction and mechanistic claims empirically traceable. The audit identifies which claim the evidence supports and which additional check is needed before a stronger interpretation is warranted.

1 INTRODUCTION

Large language models can produce fluent answers that are factually incorrect or unsupported by the available evidence. Hallucination detection seeks to identify these failures before users rely on them. Internal monitoring approaches use model activations to estimate answerability, evaluate proposed answers or rank likely errors (Kadavath et al., 2022; Azaria & Mitchell, 2023; Ji et al., 2024). These measurements also offer a way to investigate the representations associated with failure. For this reason, understanding what an internal risk score measures matters both for reliable detection and for mechanistic interpretation.

The difficulty arises when predictive evidence is used to support a stronger claim. A score that ranks errors across questions need not distinguish continuations of the same question, and features predictive of a candidate's correctness need not identify the computation responsible. Such overinterpretation can also affect answer revision: presented as verified, an auxiliary suggestion may displace an answer that was already correct. Nor does a token change after editing predictive features show that the model's use of evidence has changed. The output measure itself can mislead if swapping the tokens assigned to "supported" and "contradicted" changes whether its contrast discriminates the relation. A measurement can thus succeed at prediction while falling short of the explanation or action it is asked to support.

Prior work provides important constraints on these interpretations. Answerability differs from evaluating a supplied answer (Kadavath et al., 2022); prefix-dependent warnings change as information

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.

enters a trajectory (Huang et al., 2026); and probing research distinguishes information accessible in a representation from information used by the model (Belinkov, 2022; Hewitt & Liang, 2019). Hallucination evaluations can also depend on surface regularities, evidence conditions and label rules (Janiak et al., 2025; Chen et al., 2026). These distinctions establish why prediction alone is insufficient. The practical gap we address is how to connect them in one audit: how can a study track the same target from the observed state to predictive support and an intervention endpoint, and determine where the evidence ceases to justify the proposed claim? We ask: *what evidence licenses each step from internal prediction to mechanistic interpretation?*

We trace the score back to its information source before interpreting its use. This requires asking what the observed state can distinguish, what the readout recovers, and which representation supports that prediction. We use *risk observability*, *conditional quantitative readoutability* and *support identifiability* as operational names for these checks. They organize established distinctions around a shared target rather than introduce three independent theories. The contribution is their implementation in an evidence-chain audit, particularly the fixed-support mapping experiment that separates retained internal information from an interpretable output endpoint and a specific intervention effect.

Exact-prefix checks, candidate-conditioned readouts and support resampling first delimit the predictive evidence. A controlled nonlinear calibration tests the audit against known semantic-path supports, predictive proxies and response biases. Our main natural-model experiment then holds the source groups and selected support fixed while varying four answer codebooks and their two semantic mappings. It follows the same target across internal readout, output-endpoint discrimination and matched intervention effects.

The results separate *internal readability, output-endpoint identifiability and intervention specificity*. Gemma-2-2B-IT retains internal readout across codebooks. Under A/B, both output mappings discriminate the relation, yet neither establishes a selected-support advantage over controls. In a source-matched Qwen2.5-1.5B-Instruct stress test, weaker internal readability remains, but no evaluated output endpoint reliably discriminates the relation. These are distinct evidential limits: the extension reuses examined sources, and model-specific supports and controls differ. Figure 1 summarizes how the audit locates the unresolved transition rather than inferring a shared causal failure.

2 Related work

Risk targets and available information. Kadavath et al. (2022) distinguish answerability from evaluation of a proposed answer. Query-state probing and FacLens examine pre-generation risk (Ji et al., 2024; Wang et al., 2024), whereas statement probes, SelfCheckGPT and semantic entropy use completed text or multiple responses (Azaria & Mitchell, 2023; Manakul et al., 2023; Farquhar et al., 2024). Evaluation can change the apparent standing of a detector: Janiak et al. (2025) report strong length effects and metric sensitivity, while Chen et al. (2026) examine evidence conditions and label noise in hallucination benchmarks. We specify the target label and comparison unit because a score's reported accuracy is conditional on both.

Temporal and geometric verification. PrefixGuard studies observability and online failure warnings in agent traces (Huang et al., 2026); CLUE uses hidden-state changes across a reasoning span to verify solutions (Liang et al., 2026). Our audit concerns the interpretation of candidate-conditioned predictive structure, and layer selection at one completed-candidate position is distinct from online trajectory monitoring.

At ICLR 2026, HARP projects hidden states onto an SVD-derived reasoning subspace and reports a corresponding patch experiment; CHARM learns over token graphs combining activations and attention; Sun et al. use reasoning-depth readouts for detection and reward-shaped mitigation (Hu et al., 2026; Frasca et al., 2026; Sun et al., 2026). We do not claim to introduce internal readout or intervention; instead, we test whether prediction-selected supports are reproducible and produce target-specific effects against matched random controls.

Predictive supports and causal interpretations. Probing distinguishes accessible information from information used by the model (Belinkov, 2022; Hewitt & Liang, 2019); sparse-circuit work uses interventions to study computation (Marks et al., 2025). SAE splitting and absorption complicate

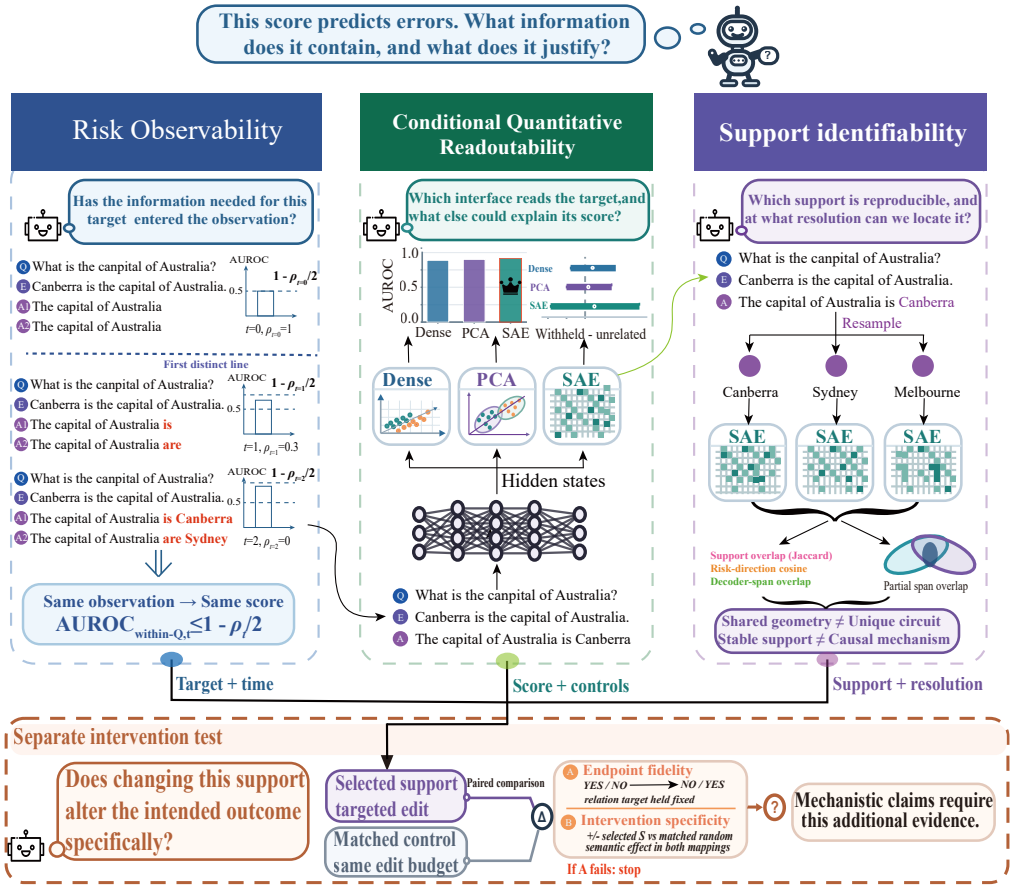

Figure 1: From an internal risk score to a stronger claim. The three properties ask whether target information is present, readable through a declared interface and reproducibly supported at a stated resolution. The separate intervention audit first checks whether the output endpoint discriminates the same target under the tested answer mapping, then compares selected-support edits with matched controls. The property panels combine distinct protocols; their arrows denote evidential checks, not a continuous generation trace or automatic progression to mechanism.

coordinate interpretations (Chanin et al., 2024), and concept subspaces need not be unique or equally well isolated across estimators (Naowarat et al., 2026; Gerasimov et al., 2026). Inference-Time Intervention, Representation Engineering and Contrastive Activation Addition show that activation changes can alter behavior (Li et al., 2023; Zou et al., 2023; Rimsky et al., 2024). Our question is narrower: does a support selected for predicting one declared target have a specific effect on that target under matched intervention controls? This separates an intervention on a predictive representation from a claim about the model's computation.

3 Tracing a Score to Its Evidential Support

3.1 A Shared Measurement Contract

An audit begins by specifying exactly which risk the score is meant to rank. Let Q denote a question, E the available evidence, and $Z_{1:t}$ a realized prefix. At layer ℓ, the activation is $H_{\ell,t} = h_{\theta,\ell}(Q, E, Z_{1:t})$, with model weights θ fixed. A condition c specifies the model, prompt, layer, token position, generation policy, evidence regime, label rule, evaluation distribution and comparison unit. The monitor accesses $O_c = \phi_c(H_{\ell,t})$ and returns $s = f_c(O_c)$. Adding evidence or a completed candidate changes c, even when the model weights stay the same.

The audit starts with a declared target Y. Expected failure across future samples, correctness of one continuation, and compatibility of a supplied answer with evidence are distinct targets. Exact match to an answer key is only an operational label, not by itself a semantic hallucination judgment. The positive class and score direction are fixed by each protocol.

The audit follows the observation, score and proposed support, $O_c \to s \to S$, while holding the target Y explicit. These arrows denote measurement dependencies, not a causal graph. An output edit requires another check: does the endpoint still discriminate Y under the answer mapping used to measure the effect?

3.2 Three questions before interpreting a score

What distinctions are available? A detector cannot distinguish outcomes that produce the same observation. *Risk observability* records whether O_c contains information about the declared target, at the intended comparison unit. Across-question risk and within-question branch discrimination must therefore be checked separately. A failed fitted probe alone cannot establish that information is absent.

What information does the interface recover? Available information need not survive a projection, sparse selection or restricted score family. *Conditional quantitative readoutability* measures held-out discrimination for a fixed interface and observation condition. Dense, PCA and SAE comparisons answer this question only when labels, candidate pools and evaluation units match. Surface controls further test what the recovered signal represents; high AUROC alone does not establish evidence use.

At what resolution can the predictive support be located? A predictive score can be realized by different coordinates or directions. *Support identifiability* asks whether the support can be located reproducibly within a declared family and metric. We distinguish coordinate overlap, decoder-span alignment and risk-direction agreement. Resampling tests the selector's reproducibility, not uniqueness over all possible supports or causal necessity. Formal definitions appear in Appendix A.1–A.2.

The order follows the proposed claim: locating a predictor presupposes a valid readout of an observable target. It is not a required sequence of computation inside the model. The checks adapt familiar information, prediction and reproducibility criteria to one measurement contract; our experiments test where evidence fails to carry over.

3.3 Two checks for the observation claim

Let $X_t = (Q, E, Z_{1:t})$ contain all inputs determining the observed state under deterministic evaluation. Draw an opposite-label pair within the same question using the declared pair weighting, and let $\rho_t = \Pr(X_t^+ = X_t^-)$.

Proposition 1 (Exact-prefix collision constraint)**.** *A deterministic state-only score satisfies* $\mathrm{AUROC}_{\mathrm{within},t} \le 1 - \rho_t/2$*. At an identical pre-generation state, within-question AUROC equals* 0.500 *whenever both labels occur.*

Proof. Colliding inputs produce equal states and scores, contributing $1/2$ to tie-aware AUROC. Each remaining pair contributes at most one. Averaging yields the bound. □

This ceiling can be loose. Its empirical value depends on the sampled trajectories; the inequality follows from the observation assumptions. It does not forbid estimating question-level risk before generation or prefix-conditioned risk after distinguishing tokens appear.

Evaluation must use the intended pairs. With N_+ positives and N_- negatives, let $W = \sum_q n_{q,+} n_{q,-}$ and $\lambda = W/(N_+ N_-)$. Splitting the standard AUROC sum gives

$$\mathrm{AUROC}_{\mathrm{pooled}} = \lambda\, \mathrm{AUROC}_{\mathrm{within,micro}} + (1-\lambda)\, \mathrm{AUROC}_{\mathrm{between}}\,. \tag{1}$$

A score constant within question contributes $1/2$ to the within term while still ranking between-question pairs. With one positive and $K-1$ negatives per question, $\lambda = 1/Q$ for Q questions, so pooled AUROC becomes dominated by between-question ordering as Q grows. This explains

why pooled success need not establish branch selection. The explicit between-question sum and the distinction from question-macro averaging are in Appendix A.

3.4 THE ADDITIONAL INTERVENTION CHECK

A reproducible predictor specifies what to edit, not what the edit means. We first evaluate the unedited output endpoint under each semantic mapping, then compare selected-support effects with matched controls. Endpoint discrimination is necessary for interpreting this contrast, not sufficient for mechanistic identification. The audit returns the supported claim and the unresolved check; answer-selection or repair benefits require separate action outcomes and cost comparisons.

4 EXPERIMENTAL DESIGN

Observation and readout. We first recompute collision ceilings from 12 archived conditions containing 38,400 trajectories: four model settings, three temperatures, 200 TriviaQA questions and 16 continuations per condition (Joshi et al., 2017). These ceilings concern realized generation prefixes. The main readout audit instead uses 160 2Wiki source questions and 500 previously examined HotpotQA questions with four supplied candidates each (Ho et al., 2020; Yang et al., 2018). All interfaces see the same 2,000 candidate-conditioned Gemma-2-2B L18 forwards, using within-question centered states, source-selected PCA and a fixed GemmaScope dictionary (Lieberum et al., 2024). States use bfloat16. NLL and entropy are single-token boundary statistics, not full-sequence likelihood or multi-sample semantic entropy. This audit is retrospective; a re-encoded supplied candidate is not the original generation trajectory.

Surface checks use a separate 120-question 2Wiki development registry: 410 of 960 candidates terminate with EOS and 550 reach the 32-token cap. Length controls diagnose a plausible shortcut, not its complete removal; generation settings appear in Appendix A.4.

Geometry and output. Twelve grouped source subsamples select supports in one dictionary. We compare coordinate Jaccard, decoder-span overlap and signed residual risk-direction cosine with equal-size active-random supports; the 66 pairwise comparisons are dependent. These tests ask how precisely a predictive support can be relocated. To test whether such a support changes an output, a separate audit ranks supports on 250 questions and applies fixed, norm-matched edits on 250 disjoint questions. Its endpoint is an evidence-verdict margin, not answer correction. For state h_q, direction v and dose ϵ, $d_q(v;\epsilon) = \frac{1}{2}[m_q(h_q + \epsilon v) - m_q(h_q - \epsilon v)]$, where m_q is the gold-option margin over wrong options. Selected-minus-random contrasts average within source before bootstrap resampling.

Frozen closure and retrospective stress tests. The known-structure calibration makes a false mechanistic attribution directly testable. A fixed two-layer nonlinear family contains two semantic-path supports, two predictive proxies, two downstream response biases and six noise blocks, with block roles permuted across instances. We freeze target, observation, support budget, thresholds, state-swap operator, controls and endpoint before evaluating 64 fresh instances. The selector sees development blocks and permitted swaps, but not role names, response weights or held-out outcomes. A separate natural-model stress test uses Gemma-2-2B-IT L18 and 106 inspected source contrasts, applying both signs of selected and norm-matched random directions at the final verdict token. This is a retrospective transfer check under a different observation protocol.

Same-source, fixed-support output-mapping audit. The natural-model audit asks whether a selected support affects the candidate–evidence relation rather than an answer token's baseline preference. We crossed two candidate identities, original versus reversed comparison direction, four single-token answer codebooks (YES/NO, A/B, X/Y and red/blue), and both support mappings per codebook while holding evidence fixed. After eligibility filtering, 20 development and 42 source- and title-disjoint test groups remained. The Gemma-2-2B-IT L18 Top-8 SAE support and eight activity-matched random supports were selected on development groups and frozen before the original YES/NO confirmation. The later multi-codebook extension reuses these already opened 42 groups and is retrospective. Qwen2.5-1.5B-Instruct uses the same 20/42 source split, but a separately frozen dense direction and coordinate-permutation controls; its support resolution and control family are not equivalent to Gemma's. For each model, 168 candidate–question conditions are crossed with

eight codebook orientations, yielding 1,344 prompt conditions. We report within-source AUROC for the frozen internal readout and the unedited two-token semantic output margin, then the selected-minus-control relation interaction from sign-balanced, norm-matched edits. Sources, not prompts or patch directions, are the independent bootstrap units. Full methods and all 16 model-by-codebook-by-orientation cells appear in Appendix A.13.

Intervals resample questions, source contrasts or simulator seeds as appropriate, conditional on recorded predictions or edits. The crossover averages within source before 10,000 bootstrap draws; repeated candidates and directions are not independent sample sizes. Protocol hashes, development records and held-out selections are archived; implementation details appear in Appendix A.

5 RESULTS

5.1 THE OBSERVATION CONSTRAINS WHICH RISK CAN BE MEASURED

Every initial-state collision ceiling is 0.500. At temperature 0.7, three-token ceilings range from 0.943 to 0.961 (Appendix Table 8). The increase marks the first opportunity to distinguish some realized continuations, not an improvement in a fitted detector. The same initial state can support question-level risk estimation while forcing a deterministic monitor to tie the future branches of that question. This is why the comparison unit must be declared before reporting an AUROC.

The candidate audit also exposed a timing error: entropy read before the final distinguishing token gave identical scores, whereas reading after the candidate entered the state gave AUROC 0.655. The change came from the observation point, not from a better classifier. It does not imply that all earlier entropy measurements are uninformative.

The full prefix plot is provided in Appendix Figure 3.

5.2 ONCE CANDIDATES ARE OBSERVED, MULTIPLE INTERFACES RETAIN SIGNAL

Supplying the candidate removes the exact-prefix collision for the comparisons made here. On the same 2,000 candidate-conditioned states, dense residuals achieve within-question AUROC 0.939 and fixed-option selection accuracy 0.858, with PCA close behind (Figure 2; Appendix Table 9). Candidate-relative correctness under the exact-match label is therefore highly readable in this retrospective setting. The weaker Top-32 SAE result shows that this sparse selection is not needed for strong readout under these conditions.

Dense exceeds PCA by 0.019 AUROC [0.001,0.036], whereas their selection difference is 0.024 [-0.008,0.058]. Ranking candidates and choosing one are already different endpoints. NLL is a final-token statistic and local entropy is not semantic entropy. The below-chance norm control does not make activation magnitude irrelevant, since reversing its score direction changes AUROC. An external Qwen-7B verifier reaches 0.996 AUROC and 0.990 selection accuracy under a different model and prompt; it is not an equal-compute internal baseline. These readouts identify useful predictive information, leaving open which representation carries it reproducibly.

5.3 PREDICTIVE COORDINATES CHANGE WHILE SOME GEOMETRY IS SHARED

For historical source-selected $K = 8$ supports, coordinate Jaccard is 0.330, decoder-span overlap 0.489 and signed risk-direction cosine 0.573; active-random supports yield 0.042 and 0.033 for the latter two (Appendix Figure 4). Thus resampling recovers some predictive geometry beyond the matched controls, even when the selected coordinates change. The three statistics refer to coordinates, decoder spans and signed directions, respectively, and cannot be substituted for one another.

On the historical audit split, the three ranking-half supports retained audit-half AUROC 0.774 [0.742,0.804], versus 0.558 [0.533,0.582] for matched active-random and 0.514 [0.489,0.538] for label-permuted supports. The selected-minus-random difference was 0.216 [0.180,0.252]. This is reproducible predictive signal conditional on the historical choice, not an output or causal claim.

The reproducible geometry is also largely explained by shared coordinates: a separate decomposition gives decoder-span overlap 0.584, of which shared coordinates account for 0.583. Removing them leaves 0.0017 (0.0008 under the original denominator), while selected and random families share no



Figure 2: Interface readout under two protocols. Panels (a,b) report within-question AUROC and top-1 selection accuracy on 2,000 supplied-candidate states from 500 Gemma-2-2B L18 questions. Panel (c) uses 184 contrasts from a separate development audit of supplied, withheld and unrelated evidence. Its changes in score ranking show evidence sensitivity, not relation specificity.

coordinates and average 0.0216. These results locate predictive information at a tested resolution without establishing uniqueness over the declared support family. A separate Qwen layer audit documents readout variation with depth (Appendix Table 5).

5.4 Frozen same-condition closure calibration

Prediction and partial localization leave a support's functional role undecided. A minimal constructed example illustrates a stable predictive proxy with zero output effect by design (Appendix A.9); it is not evidence of natural mechanism discovery.

The frozen nonlinear calibration instead tests selection on 64 fresh instances with hidden block roles. Readout, support selection and intervention share the target, observation, support budget and state-swap operator. Ground-truth roles allow direct scoring of semantic-path recovery.

AUROC-only and readout-plus-stability selected a proxy and a response bias in every held-out model (semantic precision 0.000 [0.000,0.000]; Table 1). Final-response-effect-only reached 0.500 [0.500,0.500]. Requiring a target-aligned semantic effect and interchange agreement changed the result: the intervention-qualified rule reached precision 1.000 [1.000,1.000], recall 0.969 [0.938,0.992] and exact-set recovery 0.938 [0.875,0.984], accepting no proxy or bias block. Its predeclared precision advantage over AUROC-only was 1.000 [1.000,1.000]. The calibration therefore distinguishes the evidence needed to identify the stipulated path from evidence sufficient only for prediction.

Selected supports retained semantic-interchange agreement 0.992 [0.991,0.992] with zero restoration error. Interchange agreement alone is insufficient because an unchanged proxy can agree with an unchanged endpoint; the rule also requires a nonzero target-aligned effect. It uses development interventions unavailable to predictive baselines. The gain measures what that extra evidence resolves in a known model, not equal-cost performance or a natural-LLM circuit discovery.

Table 1: Frozen calibration on 64 held-out instances of a controlled nonlinear model family. Precision measures recovery of stipulated semantic-path roles, not natural-model circuits. Brackets are 95% model-bootstrap intervals.

Frozen rule	Semantic precision	Proxy accept.	Bias accept.
AUROC-only	0.000 [0.000, 0.000]	0.500	0.500
Readout + stability	0.000 [0.000, 0.000]	0.500	0.500
Final-response effect	0.500 [0.500, 0.500]	0.000	0.500
Intervention-qualified	1.000 [1.000, 1.000]	0.000	0.000

Table 2: Same-source output-mapping audit (42 source groups per model). AUROCs and selected-minus-control relation interaction ΔI include source-bootstrap 95% intervals. Gemma and Qwen use different support and control families; all 16 cells appear in Appendix A.13.

Model	Supported code	Readout AUROC	Endpoint AUROC	ΔI
Gemma	YES	0.679 [0.601, 0.750]	0.702 [0.619, 0.786]	+0.0482 [0.0237, 0.0729]
Gemma	NO	0.696 [0.607, 0.780]	0.369 [0.292, 0.446]	-0.0412 [-0.0621, -0.0209]
Gemma	A	0.688 [0.610, 0.765]	0.643 [0.566, 0.726]	+0.0072 [-0.0017, 0.0154]
Gemma	B	0.690 [0.613, 0.768]	0.640 [0.560, 0.720]	+0.0060 [-0.0011, 0.0131]
Qwen	A	0.571 [0.506, 0.637]	0.527 [0.455, 0.598]	-0.0102 [-0.0211, 0.0012]
Qwen	B	0.625 [0.560, 0.690]	0.530 [0.452, 0.610]	-0.0007 [-0.0097, 0.0081]

Earlier transfer and polarity checks motivate fixing the target, support and endpoint in the following audit; their separate protocols and results appear in Appendix A.7.

5.5 Internal readability, output endpoint and intervention effect separate under answer remapping

The original source-isolated Gemma confirmation held the observation point, selected SAE support and edit procedure fixed. Development out-of-fold AUROC was 0.645 [0.525,0.766], with bootstrap support Jaccard 0.204. On 42 source-disjoint confirmation groups, within-source internal-readout AUROC was 0.692 [0.613,0.769]. The subsequent multi-codebook audit retained this support and these sources. Across all eight answer-codebook orientations, the frozen internal readout remained above chance, with AUROC 0.649–0.696 (Table 2; Appendix A.13). This is a readout-transfer result, not an output or intervention result.

The unedited output margin was a separate measurement. Under YES=supported, Gemma's output-endpoint AUROC was 0.702 [0.619,0.786], but it fell to 0.369 [0.292,0.446] when NO=supported, even though the internal readout remained 0.696 [0.607,0.780]. Thus the reversed YES/NO output contrast failed to discriminate the relation in its declared direction. The A/B codebook provided a more revealing control: both mappings retained output discrimination, at 0.643 [0.566,0.726] and 0.640 [0.560,0.720]. Endpoint discrimination therefore can be present even when the intervention claim remains open. These AUROCs concern a specified pair of next-token logits; they do not measure free-form instruction following or complete semantic fidelity.

With the Gemma support still fixed, the YES/NO selected-minus-random relation interaction changed from $+0.0482$ [0.0237,0.0729] to -0.0412 [-0.0621,-0.0209] under the reverse mapping, a paired difference of $+0.0894$ [0.0487,0.1316]. Because the reversed mapping failed the endpoint check, this sign reversal is not evidence that a relation mechanism itself reversed. The raw YES-token margin shifted in the same direction under both mappings. Under A/B, where both unedited endpoints discriminated the relation, selected-minus-random interactions were $+0.0072$ [-0.0017,0.0154] and $+0.0060$ [-0.0011,0.0131]. Neither interval established a selected-support advantage. The red/blue codebook produced one nominal positive orientation, but its paired reverse orientation did not establish the same effect; no codebook was selected after viewing the test outcomes to certify mapping invariance.

The source-matched Qwen audit stopped earlier. Its separately frozen dense readout gave within-source AUROC 0.571–0.625 across the eight orientations, whereas unedited output-endpoint AUROCs ranged from 0.491 to 0.545 and every corresponding 95% interval included 0.5. Selected-minus-permuted-direction interaction intervals also included zero. Qwen's weak internal ranking cannot license an intervention conclusion when none of the evaluated output contrasts is reliably relation-discriminative. This does not replicate Gemma's particular mapping-dependent intervention pattern, nor does it show that Qwen lacks a usable endpoint under every possible prompt or codebook.

Consequences for use. The distinction between a risk indication and a verified correction also matters when a model revises an answer. In a separate 120-question 2Wiki birth-date task, a scripted but unsupported claim that an auxiliary suggestion had been internally verified left a correct prior answer intact in 33/120 cases, versus 83/120 with a scope-limited note. The latter also produced fewer beneficial corrections of wrong priors (88/120 versus 111/120). Averaging the original and a post-result reversed evidence order, its balanced advantage over direct year comparison was only $+2.1$ percentage points $[-1.9, 5.8]$. This tests the effect of a diagnostic message, not an audit-gated repair policy (Appendix A.6). A separate natural-candidate record likewise establishes no consistent selection advantage over fixed interfaces (Appendix A.5).

6 Discussion

The mapping audit adds a concrete measurement failure to the familiar warning that a probe is not a mechanism. Internal information can persist while an output contrast loses its semantic discrimination. A raw answer-token preference can then yield sizeable patch effects, and reversing the semantic mapping can reverse their interpretation without changing the edited support. Gemma's A/B result further separates a discriminative endpoint from evidence of selected-support specificity. Qwen stops earlier at the endpoint check; it does not replicate Gemma's intervention pattern. A zero-crossing contrast interval indicates uncertainty, not absence of causal influence.

The three properties maintain target continuity before an output edit is interpreted; they do not define a new taxonomy of mechanisms. Known-path calibration shows what additional intervention evidence resolves when semantic outcomes are available. Natural-model roles remain unknown, and a same-protocol positive control is still needed to establish the assay's sensitivity to a genuinely relation-specific effect. The cost of overinterpretation also appears in answer revision: a false verification claim can displace a correct prior answer. Scope-limited wording curbs this error, but it also prevents some warranted corrections, and its balanced accuracy does not reliably exceed direct evidence comparison. This result motivates an audit-gated repair policy without validating one. Such a policy still needs independent, cost-matched evaluation of final quality and side effects.

7 Limitations

The natural-model evidence concerns small instruction-tuned models and supplied-answer tasks. The original Gemma YES/NO crossover was prospective for model outcomes, although source records and annotations were screened beforehand; the extension to four codebooks reuses opened sources and is retrospective. Qwen shares the historical source split but uses a different support family and control construction, so it is a stress test, not a matched replication of Gemma's sparse intervention. Source-bootstrap intervals do not account for codebook selection after prior inspection, multiple comparisons or support refitting. The release-year comparison measures neither free-generation hallucination incidence nor mitigation. A first-token output AUROC does not establish unconstrained answer semantics or instruction compliance, and the same-protocol natural-model positive control remains unfinished. The answer-revision test uses scripted notes rather than model-derived audit outputs; its order checks are post-result and cannot estimate natural misdiagnosis rates. The controlled model has known roles and gives the intervention-qualified rule extra development evidence; its support-recovery advantage is not an equal-cost natural-model result. No unique, necessary or sufficient natural circuit or repair has been established. Appendix A.16 gives further limits.

8 Conclusion

An internal score can be valuable for prediction while leaving the proposed mechanism unverified. The three-property audit tests what is observable, readable and reproducibly supported; a separate intervention audit checks whether the output endpoint discriminates the target before interpreting matched edits. In natural-model tests, the empirical chain stops at different gates across models and mappings. This locates unsupported claim upgrades rather than identifying a causal hallucination circuit. Cost-matched action outcomes from an actual audit-gated policy remain necessary before claiming better answer selection or repair.

AI USE STATEMENT

Generative AI tools assisted with developing and refining the audit framework, hypotheses, mathematical exposition and experimental designs; implementing and debugging analysis code; preparing candidate and annotation materials; interpreting results; literature search; figure preparation; and drafting and revising the manuscript. Model-assisted annotation was used in parts of the research workflow and is distinct from human adjudication. AI-generated suggestions and labels are not independent evidence of a mechanistic claim. The authors are responsible for the final text, code, data, citations and scientific interpretations.

REPRODUCIBILITY STATEMENT

The analyses combine archived trajectories, retrospective audits, an originally source-isolated Gemma model-outcome evaluation, a source-matched Qwen stress test, a seed-held-out calibration and a source-ID-screened answer-revision test. Splits, hashes, frozen model-specific supports and estimands are documented in the appendices and experiment files. The four-codebook extension reuses opened source groups; the answer-revision order checks were designed after the initial results. Neither is described as untouched-source confirmation.

REFERENCES

Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 967–976, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.68. URL https://aclanthology.org/2023.findings-emnlp.68/.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422. URL https://aclanthology.org/2022.cl-1.7/.

David Chanin, James Wilken-Smith, Tomáš Dulka, Hardik Bhatnagar, Satvik Golechha, and Joseph Bloom. A is for Absorption: Studying Feature Splitting and Absorption in Sparse Autoencoders. *arXiv preprint arXiv:2409.14507*, 2024. doi: 10.48550/arXiv.2409.14507. URL https://arxiv.org/abs/2409.14507.

Wenbo Chen, Veena Padmanabhan, Tootiya Giyahchi, Elaine Wong, and Leman Akoglu. Rethinking evaluation for LLM hallucination detection: A desiderata, a new RAG-based benchmark, new insights. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14912–14931, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.680. URL https://aclanthology.org/2026.acl-long.680/.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, 2024. doi: 10.1038/s41586-024-07421-0. URL https://www.nature.com/articles/s41586-024-07421-0.

Fabrizio Frasca, Guy Bar-Shalom, Yftah Ziser, and Haggai Maron. Neural message-passing on attention graphs for hallucination detection. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/014898b821e3ce966ae1db68cb8d9966-Abstract-Conference.html.

Gleb Gerasimov, Timofei Rusalev, Nikita Balagansky, Daniil Laptev, Vadim Kurochkin, and Daniil Gavrilov. Unstable Features, Reproducible Subspaces: Understanding Seed Dependence in Sparse Autoencoders. *arXiv preprint arXiv:2606.12138*, 2026. doi: 10.48550/arXiv.2606.12138. URL https://arxiv.org/abs/2606.12138.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

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*, pp. 6609–6625, Barcelona, Spain (Online), 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.580. URL https://aclanthology.org/2020.coling-main.580/.

Junjie Hu, Gang Tu, ShengYu Cheng, Jinxin Li, Jinting Wang, Rui Chen, Zhilong Zhou, and Dongbo Shan. HARP: Hallucination detection via reasoning subspace projection. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/1e58b1bf9f218fcd19e4539e982752a5-Abstract-Conference.html.

Xinmiao Huang, Jinwei Hu, Rajarshi Roy, Changshun Wu, Yi Dong, and Xiaowei Huang. PrefixGuard: From LLM-Agent Traces to Online Failure-Warning Monitors. *arXiv preprint arXiv:2605.06455*, 2026. doi: 10.48550/arXiv.2605.06455. URL https://arxiv.org/abs/2605.06455.

Denis Janiak, Jakub Binkowski, Albert Sawczyn, Bogdan Gabrys, Ravid Shwartz-Ziv, and Tomasz Jan Kajdanowicz. The illusion of progress: Re-evaluating hallucination detection in LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 34728–34745. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1761. URL https://aclanthology.org/2025.emnlp-main.1761/.

Ziwei Ji, Delong Chen, Etsuko Ishii, Samuel Cahyawijaya, Yejin Bang, Bryan Wilie, and Pascale Fung. LLM internal states reveal hallucination risk faced with a query. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 88–104, Miami, Florida, US, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.blackboxnlp-1.6. URL https://aclanthology.org/2024.blackboxnlp-1.6/.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, Vancouver, Canada, 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147/.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language Models (Mostly) Know What They Know. *arXiv preprint arXiv:2207.05221*, 2022. doi: 10.48550/arXiv.2207.05221. URL https://arxiv.org/abs/2207.05221.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems 36*, 2023. doi: 10.52202/075280-1797. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/81b8390039b7302c909cb769f8b6cd93-Abstract-Conference.html.

Zhenwen Liang, Ruosen Li, Yujun Zhou, Linfeng Song, Dian Yu, Xinya Du, Haitao Mi, and Dong Yu. Your reasoning model is secretly a reward model - optimization-free verification from experience. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 17358–17372, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.788. URL https://aclanthology.org/2026.acl-long.788/.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma Scope: Open Sparse Autoencoders Everywhere All At Once on Gemma 2. *arXiv preprint arXiv:2408.05147*, 2024. doi: 10.48550/arXiv.2408.05147. URL https://arxiv.org/abs/2408.05147.

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017, Singapore,

2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.557. URL https://aclanthology.org/2023.emnlp-main.557/.

Samuel Marks, Can Rager, Eric J. Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse Feature Circuits: Discovering and Editing Interpretable Causal Graphs in Language Models. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/3ba4d47a83e498c2b1a0868cba20f6de-Abstract-Conference.html.

Burin Naowarat, Hao Tang, and Sharon Goldwater. A framework for analyzing concept representations in neural models. In *Proceedings of the 30th Conference on Computational Natural Language Learning*, pp. 574–587. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.conll-main.34. URL https://aclanthology.org/2026.conll-main.34/.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

Zhongxiang Sun, Qipeng Wang, Haoyu Wang, Xiao Zhang, and Jun Xu. Mechanistic detection and mitigation of hallucination in large reasoning models. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/0271e4dc4e81459121a07b9634d5e79d-Abstract-Conference.html.

Yanling Wang, Haoyang Li, Hao Zou, Jing Zhang, Xinlei He, Qi Li, and Ke Xu. FacLens: Transferable Probe for Foreseeing Non-Factuality in Fact-Seeking Question Answering of Large Language Models. *arXiv preprint arXiv:2406.05328*, 2024. doi: 10.48550/arXiv.2406.05328. URL https://arxiv.org/abs/2406.05328.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William 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*, pp. 2369–2380, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023. URL https://arxiv.org/abs/2310.01405.

A Additional methods and provenance

A.1 Observation and readout definitions

For a fixed condition c and binary label Y, risk is observable through O_c when

$$\Pr[\Pr(Y = 1 \mid O_c, c) \neq \Pr(Y = 1 \mid c)] > 0.$$

For within-question discrimination, restrict to $\mathcal{Q}_{\mathrm{nd}}(c) = \{q : n_{q,+} > 0 \text{ and } n_{q,-} > 0\}$ and condition both probabilities on $Q = q$. A single-class question cannot establish discrimination. Under the usual regularity conditions, this is equivalent to positive conditional mutual information on the restricted population.

For a fixed interface ϕ_c, score family $\mathcal{F}_c$, score direction and threshold $\tau > 1/2$, the target is quantitatively readable when

$$\sup_{f \in \mathcal{F}_c} \mathrm{AUROC}_{D_c}(f(O_c), Y) \geq \tau.$$

An empirical claim requires training/development selection and held-out grouped evaluation with a declared aggregation and uncertainty procedure. This condition does not guarantee that finite training finds the best score. Historical results without jointly fixed thresholds are estimates, not retrospectively certified passes.

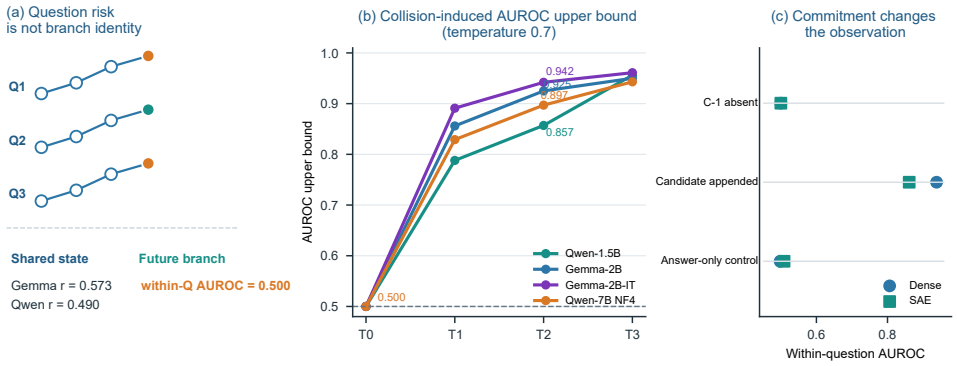

Figure 3: Risk observability depends on the observation point. Equal prefixes force tie-aware within-question AUROC to 0.500. Distinguishing tokens relax the collision-induced ceiling, which is a bound rather than a detector score.

A.2 Formal support identifiability

Definition 1 (Support identifiability within an audit family)**.** *Assume a usable readout under the same target and condition. Let $\mathcal{S}_c$ be a nonempty declared support family and $R_c(S)$ one minus its best admissible AUROC. For tolerance ϵ, define*

$$\mathcal{E}_{\epsilon,c} = \{S \in \mathcal{S}_c : R_c(S) \le \inf_{S' \in \mathcal{S}_c} R_c(S') + \epsilon\}.$$

Localization to $S^\star$ at resolution η requires $\sup_{S \in \mathcal{E}_{\epsilon,c}} d(S, S^\star) \le \eta$, together with a selection procedure returning supports within that resolution with probability at least $1-\alpha$ under the declared resampling scheme.

Let $\widehat{S}_{\mathcal{D}}$ denote the support returned by the declared selection procedure on $\mathcal{D} \sim \mathcal{R}_c$. When a reference $S^\star$ is available, a finite-sample estimate of the resolution-η selection probability is

$$\widehat{p}_{B,\eta} = \frac{1}{B} \sum_{b=1}^{B} \mathbf{1}\{d(\widehat{S}_{\mathcal{D}_b}, S^\star) \le \eta\}.$$

A.3 Collision ceilings and conditional information

If G_t indexes exact complete-input classes, then H_t is deterministic given G_t under the stated evaluation assumptions, so $I(Y; H_t \mid G_t) = 0$. This does not imply $I(Y; H_t) = 0$ across questions. The empirical bound counts opposite-label prefix collisions, using pooled within-question pairs for its micro weighting. Question-macro weighting gives a different average. At later prefixes the simple ceiling may be loose because two distinct classes can each contain both labels.

For $0 < W < N_+ N_-$, the between-question term in Eq. 1 is

$$\mathrm{AUROC}_{\mathrm{between}} = \frac{1}{N_+ N_- - W} \sum_{q \neq q'} \sum_{i \in \mathrm{Pos}_q} \sum_{j \in \mathrm{Neg}_{q'}} \left[\mathbf{1}(s_i > s_j) + \tfrac{1}{2}\mathbf{1}(s_i = s_j)\right].$$

It excludes within-question pairs; a question-macro average uses different weights and cannot replace the micro term in Eq. 1.

For one question, let class g contain a_g positives and b_g negatives, with local totals $N_{q,+} = \sum_g a_g$ and $N_{q,-} = \sum_g b_g$. Arbitrary hindsight class scores can attain

$$\frac{\frac{1}{2}\sum_g a_g b_g + \sum_{g<h} \max(a_g b_h, a_h b_g)}{N_{q,+} N_{q,-}}$$

by ordering empirical class-positive fractions. This is the question-local AUROC; a dataset-level micro quantity requires summing the corresponding numerator over questions and dividing by $\sum_q N_{q,+} N_{q,-}$. The finite-sample expression may lie below the collision-only ceiling. It uses the observed labels and must not be reported as a learned detector's generalization performance.

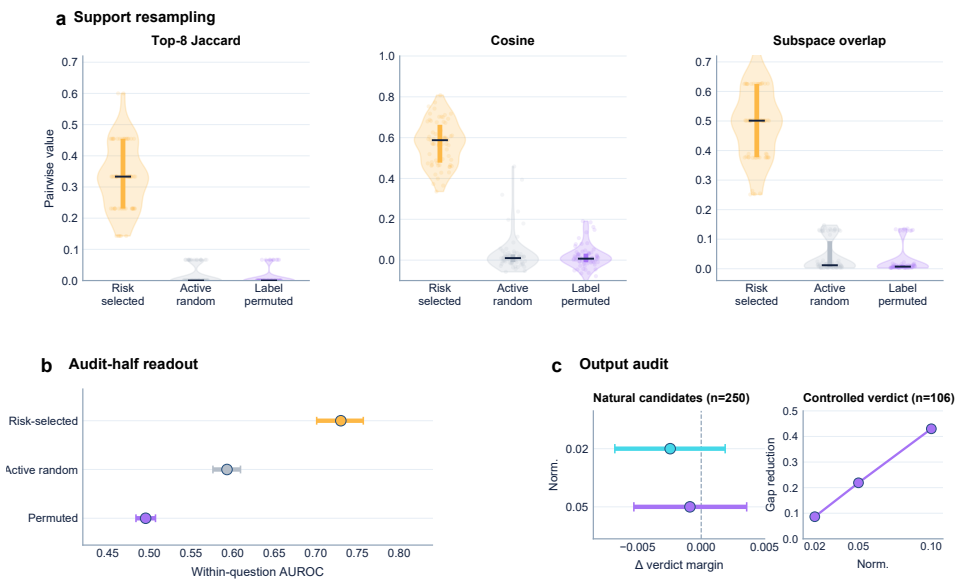

Figure 4: Predictive support and retrospective output checks. Resampling retains alignment above matched random supports, although shared coordinates explain much of the decoder-span overlap. The verdict curve uses earlier label-directed patch signs; the sign-balanced specificity audit is reported separately.

A.4 Readout and geometry

The separate 120-question 2Wiki surface-control registry uses Gemma-2-2B-IT, temperature 1.0, top-p 0.95, a one-to-32-token budget and deterministic seeds. Of 960 candidate slots, 410 terminate with EOS and 550 reach the cap. The common-interface audit fits on 160 source questions (640 candidates). PCA rank is selected among 8,16,32,64 on source data; the reported rank is 64. Dense and PCA operate on within-question centered states. The SAE interface uses Top-32 source-selected features of the fixed GemmaScope L18 width-16k dictionary; for each source fit, coordinates are ranked by the absolute standardized logistic coefficient, not by activation frequency. The target consists of four supplied candidates for each of 500 previously inspected questions. The Gemma-2-2B-IT states are extracted in bfloat16. No additional stochastic sampling occurs in that protocol. The final-candidate NLL and post-candidate entropy rows are local one-token statistics; full-sequence length-normalized likelihood and multi-sample semantic entropy are outside this table's scope.

For equal-rank decoder spans with orthogonal projectors P_i, P_j, overlap is $\mathrm{tr}(P_i P_j)/K$. Direction cosine uses decoder directions weighted by the fitted readout. At $K = 8$, the coordinate, span and direction summaries are 0.329951,0.488943,0.572597. The corresponding active-random span and direction summaries are 0.041596 and 0.032887. Twelve source subsamples yield 66 dependent pairs. The separate Qwen candidate-conditioned audit uses 66 development and 506 confirmation questions and must not be pooled with this supplied-candidate experiment.

A.5 Natural-candidate selection and fallback

This separate local record contains 506 HotpotQA questions and eight candidates per question generated by Gemma-2-2B-IT with the chat template, temperature 1.5, top-p 0.95 and a 32-new-token cap. The same candidate pool is re-encoded by Gemma-2-2B-IT and Qwen2.5-1.5B-Instruct, both in bfloat16, at their respective L18 residual interfaces. Development uses 66 questions and 528 candidates. The outcome label is conservative exact match to the private official answer. The archive records 4,030 of 4,048 candidate states; all 18 questions with a missing state remain in the denominator and use the first candidate. Questions without a correct candidate are retained. These

choices make the record a direct but retrospective decision audit, rather than a continuation of the supplied-option protocol.

Development readouts use standard scaling and class-balanced logistic regression with $C = 1$, liblinear and at most 3,000 iterations. PCA retains up to 32 training-supported components. Five-fold GroupKFold produces question-held-out development scores. The relation packet uses StratifiedGroupKFold with seeds 20260914 and 20260915. Models fit only on supplied-evidence pairs from other groups and score all four conditions for the held-out groups.

The operational relation criterion requires a positive supplied AUROC minus the mean control AUROC, and a supplied supported-minus-contradicted score margin above each control margin, in both seeds. If none qualifies, the rule uses L18 Dense. These are operational development criteria rather than calibrated hypothesis tests. In particular, PCA may retain almost all training variation in the small relation folds, so agreement with Dense is not an independent replication across representations.

Table 3: Natural-candidate exact-match selection (%) on the same 506 questions and eight Gemma-generated candidates per question. Missing-state questions are retained. The relation rule is a limited relation audit with a Dense fallback, not a full three-property controller.

Strategy	Gemma readout	Qwen readout
First candidate	4.15	4.15
Uniform random (expected)	3.58	3.58
Fixed L18 Dense	6.97	12.65
Fixed L18 PCA-32 / AUROC choice	1.19	17.39
Relation rule with Dense fallback	6.97	12.65
Candidate-length control	13.54	13.54
Token-count control	13.03	13.06
Candidate oracle	20.75	20.75

On Gemma, the relation rule uses Dense while development AUROC selects PCA. Their selection rates are 6.97% and 1.19%, a paired difference of $+5.78$ percentage points [3.93,7.78]. The relation rule therefore equals fixed Dense, and both length and token-count controls are stronger in this record. The relation gate ties Dense and PCA on its development criterion, so the choice of Dense is an alphabetical fallback; this contrast is not an incremental full-framework gain. Resolving that tie in favor of PCA reproduces the AUROC-only action. On Qwen states for the same candidates, the ordering reverses: PCA reaches 17.39%, while Dense and the relation fallback reach 12.65%. The rule-minus-PCA difference is -4.74 points [-6.92,-2.77].

All 506 natural-candidate questions enter the action denominator; the candidate oracle is 20.75%. Eighteen questions have an unrecorded state and select their first candidate. Otherwise tied top scores receive mean label accuracy, corresponding to expected uniform tie-breaking. Stored deterministic tie indices must not be confused with this expected metric. Length/token residualization fitted on development data reduces Gemma Dense to 4.36%, only $+0.78$ points above random [-0.72,2.38]. The intervals use 20,000 question-bootstrap resamples of the fixed predictions. They do not include the uncertainty of retraining, selecting interfaces, or repeatedly revising the research protocol.

A.6 Controlled answer revision under diagnostic wording

The mechanism audit does not itself select a replacement answer. We therefore tested a narrower downstream question: what happens when an auxiliary suggestion is described as verified without a check that could warrant that description? This experiment manipulates a message about diagnosis, not an actual monitor output. It cannot estimate the prevalence of such messages in deployed systems.

We used the locally cached 2Wiki validation split. Eligible questions directly compared which of two named people was born earlier or later and had exactly two source-derived birth-date triples with unambiguous years. We required the answer implied by the years to match the dataset answer. Of 522 structurally eligible questions, 26 shared source IDs with specified previously opened local packets or registries and were excluded. Hash-ranked sampling selected 30 questions from each earlier/later by correct-person-position stratum, yielding 120 source questions. This screening checks

the named source IDs, not all entity or document overlap with prior research. Each prompt showed both birth-date triples, a scripted prior assistant answer and a conflicting auxiliary suggestion. We crossed correct prior/wrong suggestion with wrong prior/correct suggestion. The instruction-tuned Gemma-2-2B model generated one answer per condition with greedy decoding and a 48-token cap, without retries. Each of five message arms was applied to both conditions, for 1,200 initial generations.

The five arms left the question, evidence and candidate identities unchanged. The no-note arm only supplied the auxiliary suggestion. The false-verification arm said an internal check had verified the suggestion and the prior answer should be replaced; no such check had been performed. Generic caution asked the model to compare both answers with the evidence. The scope-limited note said that an internal risk signal could indicate possible error but did not verify the suggestion or certify a repair, and asked the model to compare it with the evidence. A direct-year-check arm instructed the model to extract and compare the two birth years while treating the suggestion as unverified. The first four wordings were carried over unchanged from a 19-question development pilot. All arms requested a candidate name or *UNCERTAIN* and a categorical confidence line. Nonconforming answers stayed in the denominator. Prompt lengths differed; the total prompt-token counts over the original and evidence-swapped runs were 70,574 (no note), 83,054 (false verification), 84,974 (generic caution), 91,214 (scope-limited) and 92,174 (year check). These are not equal-cost strategy comparisons.

Table 4: Final answers consistent with the supplied birth-date evidence in the 120-question wording test. Each entry has denominator 120. PC/W: prior answer correct, suggestion wrong; PW/C: prior answer wrong, suggestion correct. The reversed-evidence conditions were run after the initial position effect was observed, not as a second independent confirmation.

Message arm	Original PC/W	Original PW/C	Reversed PC/W	Reversed PW/C
No note	67	90	50	92
False verification	33	111	37	104
Generic caution	55	81	51	90
Scope-limited	83	88	60	77
Direct year check	81	70	75	72

The predeclared initial contrast concerned harmful replacement when the prior answer was correct and the suggestion wrong. Scope-limited wording exceeded false verification by 41.7 percentage points [33.3,50.8], but it lost 19.2 points [-26.7,-12.5] when the prior was wrong and the suggestion correct. The two conflict cells' equal-weight difference was +11.3 points [5.4,17.1] under the original order. Intervals use 10,000 source-question paired bootstrap resamples of fixed replies; they do not account for prompt design, model choice or re-selection of questions.

Inspection of that result revealed that the correct person's original position strongly modified performance. We then ran two post-result crossover diagnostics, each with another 1,200 generations on the same sources. Swapping only the two names in the question left the position interaction largely intact. Swapping only the two evidence lines reduced the scope-limited note's two-order balanced advantage over direct year checking to +2.1 points [-1.9,5.8], and over no note to +1.9 [-1.0,4.8]. Against false verification it retained +30.4 points [24.6,36.3] in the correct-prior/wrong-suggestion cell, but lost 20.8 points [-25.8,-15.8] in the beneficial-correction cell; the equal-weight difference was +4.8 [0.8,8.8]. This is evidence of a trade-off under the constructed 50:50 conflict mixture, not a natural-frequency estimate of total harm. The post-result checks diagnose sensitivity and do not constitute independent replications.

The earlier 19-question development pilot showed the same directional trade-off without a resolved overall gain: scope-limited minus false verification was +5.3 points [-13.2,+23.7] on the equal-weight endpoint. In a separate 19-question evidence-redaction pilot, precise birth years were withheld, although two items retained decade-level cues. The scope-limited note produced zero explicit *UNCERTAIN* answers in 38 conflict prompts, versus 7/38 under generic caution. Reduced self-reported confidence therefore did not amount to evidence-appropriate abstention. None of these wording tests used a model-derived three-property diagnosis or an online audit gate, and none established a cost-matched improvement in natural answer selection or repair.

A.7 EARLIER TRANSFER AND POLARITY CHECKS

Three supports selected on a ranking partition have AUROCs 0.811, 0.801 and 0.768. On a disjoint 250-question audit partition, mean antisymmetric verdict-margin effects are 0.00394 for selected and 0.00311 for matched active-random directions, a difference of 0.00083 $[-0.00556, 0.00739]$. This is a continuous margin effect, not label flips. Because the endpoint changes, it tests transfer rather than closing the same prediction-to-intervention chain.

The earlier polarity experiment used relation labels to choose patch signs, allowing a generic affirmative shift to resemble a relation effect. A retrospective crossover therefore applies both signs to every prompt (Appendix A.11). Across 106 source contrasts, selected directions produce a common shift of 0.01902 [0.01592,0.02221], while the selected-minus-random relation interaction is 0.00285 $[-0.00452,0.01051]$ (Appendix Table 10). The reversed mapping fails baseline fidelity. This establishes an output shift without establishing relation specificity; it does not exclude an additional causal contribution.

A.8 RETROSPECTIVE QWEN LAYER SELECTION

The layer grid is 6,12,18,24,27 in the implementation's zero-based Transformer block indexing. Every capture uses the last token after the complete candidate; the index is computational depth, not decoding time. Table 5 reports development performance. The selected pair is L18 PCA, with only a small development advantage over L18 Dense. No statistical superiority of that development difference is claimed. The relation criterion fails for every combination across its two seeds. On the same confirmation record, the fixed-layer choices are 9.88% (L6 Dense), 14.43% (L12 Dense), 12.65% (L18 Dense), 11.46% (L24 Dense), 9.29% (L27 Dense), and 17.39% (L18 PCA-32); the full-framework rule falls back to L18 Dense at 12.65%. Thus broadening the layer search documents strong layer dependence but does not establish incremental relation-gated action utility or a full-framework advantage. The experiment is an offline diagnostic and does not implement per-question routing or token-by-token monitoring.

Table 5: Qwen development OOF within-question AUROC for offline layer/interface selection. The target set had already been examined before this extension.

Block index	Dense	PCA-32
6	0.7923	0.7871
12	0.8125	0.8028
18	0.8688	0.8691
24	0.8329	0.8197
27	0.8327	0.8298

Table 6: Qwen2.5-1.5B-Instruct candidate-conditioned confirmation audit. The layer/interface choice is made on 66 development questions and evaluated on 506 questions; all values use the exact-match endpoint.

Strategy	Selection accuracy (%)
Fixed L6 Dense	9.88
Fixed L12 Dense	14.43
Fixed L18 Dense	12.65
Fixed L24 Dense	11.46
Fixed L27 Dense	9.29
Fixed L18 PCA-32 / AUROC choice	17.39
Full framework (L18 Dense fallback)	12.65
Candidate-length control	13.54
Token-count control	13.06

A.9 GROUND-TRUTH CAUSAL–PROXY CALIBRATION

For each of 20 parameter seeds, draw unit vectors $u_C, u_P \in \mathbb{R}^8$. Independently generate 700 development and 900 test instances. Let y be uniform on $\{-1, 1\}$ and let $\tilde{y} = y$ with probability 0.91, otherwise $\tilde{y} = -y$. The state and output are

$$C = 1.35 y u_C + \epsilon_C, \quad P = 1.35 \tilde{y} u_P + \epsilon_P, \quad H = (C, P, N), \quad M = 1.25 C^\top u_C + \epsilon_M,$$

where $\epsilon_C, \epsilon_P \sim \mathcal{N}(0, 0.9^2 I)$, $N \sim \mathcal{N}(0, I_{48})$, and $\epsilon_M \sim \mathcal{N}(0, 0.55^2)$ are independent. Logistic probes are fitted separately to the known C and P blocks with regularization parameter 0.4. Forty development bootstrap fits per seed yield absolute coefficient cosines relative to the full-development direction. This is direction repeatability within known supports, not inferred support identifiability.

The intervention calculation sets $M_C = M - 0.9(1.25 C^\top u_C)$ and $M_P = M_R = M$. Signed loss is $y(M - M_\cdot)$. Restoration adds back the same causal term; its mean absolute residual is 1.32×10^{-18}. The proxy and random effects are structurally zero, and restoration is algebraic. This is a ground-truth removal of a known term, not a claim that a matched-norm hidden-state edit through an independently learned network has identified a natural circuit. The reported interval uses 10,000 seed-bootstrap draws. These numbers calibrate the specificity of the audit under the stated equation; they must not be interpreted as a causal discovery success rate for a natural language model.

A.10 FROZEN SAME-CONDITION CLOSURE CALIBRATION

The closure protocol freezes 64 evaluation seeds before their networks are generated. Each network has 12 role-permuted blocks: two semantic-path supports, two label-predictive proxies, a generic and a weak downstream response bias, and six noise blocks. The declared target is the semantic verdict determined by the two semantic blocks. The final response adds the downstream bias, so it is a distinct endpoint. For every frozen rule, the observation is one candidate block, the support budget is two blocks, and intervention swaps the selected block with a block from an opposite-label source state. Development and held-out states contain 900 and 1,400 instances per model, respectively.

The selector receives shuffled IDs, development blocks, semantic labels and, for the intervention-qualified rule, development swap outcomes. It is denied role names, response weights and all held-out outcomes. The intervention-qualified rule requires AUROC at least 0.70, bootstrap direction stability at least 0.82, a lower 95% bound above 0.025 for the semantic-aligned swap effect, and semantic interchange agreement at least 0.985. Held-out evaluation restores role labels only after selection. The frozen protocol, implementation hashes, development metrics, 320 held-out records and result hashes are archived with the code. A separate deterministic replay reproduced byte-identical CSV and JSON result files. This is procedural seed-held-out confirmation rather than an externally preregistered or independently executed study.

A.11 CROSSOVER IMPLEMENTATION AND CODEBOOK CALIBRATION

The crossover uses 106 historical source contrasts with a fixed candidate and evidence but opposite question relations. It imports selected and active-random $K = 8$ residual directions for historical replicates 2, 6, and 3. Directions are normalized to a common edit norm within each source, equal to 0.02 times the root-mean-square norm of that source's two archived states. The archive predates the explicit answer-mapping prompts; the dose is therefore not 2% of the newly prompted state's measured norm. Each direction is added at the final non-padding token of zero-based block 18 of Gemma-2-2B-IT. The score is the next-token log-probability difference between the mapped support and contradiction answer words. The model uses the archived inference configuration, including its numerical precision; these are conditional forward-pass measurements rather than natural answer repairs.

For a mapped support margin m_r and relation $r \in \{s, c\}$, the crossover contrast is $d_r(v) = \frac{1}{2}\{m_r(+v) - m_r(-v)\}$. We report the relation interaction $I(v) = d_s(v) - d_c(v)$ separately from the common signed shift $B(v) = \frac{1}{2}\{d_s(v) + d_c(v)\}$. This separation prevents a label-directed polarity rule from turning a generic affirmative-response bias into an apparent relation effect.

Every relation receives both patch signs under both answer mappings. There are 424 unpatched prompt evaluations and 5,088 patched evaluations, yielding 1,272 source–mapping–family–direction

effects before within-source averaging. None of these repetitions increases the number of independent source contrasts beyond 106. Source-bootstrap intervals use 10,000 draws; three fitted directions are not three independently trained models. The lower bound for the reversed-mapping selected-minus-random interaction is approximately 4.49×10^{-9} before rounding, but baseline fidelity fails under that mapping. We do not use that borderline interval as mechanistic confirmation.

The codebook follow-up checks a fixed ordering of A/B, 1/2, X/Y, and red/blue, retaining only pairs whose two response words each occupy one continuation token. The recorded eligible pairs are A/B, X/Y, and red/blue. A deterministic source hash assigns 64 sources to development and 42 to the held-out calculation. The selection function chooses the first codebook with positive lower 95% baseline-gap bounds in both development orientations. X/Y satisfies that rule: gaps are 0.06738 [0.01660,0.11914] and 0.10742 [0.03516,0.18457]. The remaining-source gaps are 0.03869 [-0.05655,0.13690] and 0.15625 [0.06548,0.24851]. Failure to establish a positive gap in both orientations is not proof that the gap is zero.

All codebooks' held-out scores were computed in the same execution before the selection summary was written, although the selection function uses development values only. The data had also been inspected historically. We therefore report a development-selected encoding with a retrospective held-out check, not a prospectively frozen confirmation. The archived crossover decision rule and the later endpoint-fidelity assessment are distinct; their combination is not an externally preregistered gate. No intervention results under the X/Y codebook are reported.

A.12 Prospective natural-model polarity crossover

The experiment crosses candidate identity (official or alternate), comparison direction (original or reversed), and response mapping (YES or NO denotes support). Evidence remains unchanged within each source group. Eligibility is determined mechanically from title-anchored evidence: each of the two entities must have one explicit year, the years must differ, the candidate must match one title, and the official answer must agree with the year ordering. Development also requires agreement by both reviewers that the original relation is supported; reviewer-proposed edits are not used. The 20 development and 42 confirmation groups, as well as their supporting titles, are disjoint. The source pool was screened before model outcomes were collected, so the prospective designation applies to the model-state, score, and intervention evaluation rather than source acquisition.

The SAE readout uses five-fold GroupKFold by source group. Within each training fold, activations are standardized and features are ranked by the absolute supported-minus-contradicted mean difference; held-out rows receive the signed top-eight score. The development support is refit on all development groups and frozen before confirmation access. Reproducibility is summarized by feature-set Jaccard over 200 source-group bootstrap fits. For the intervention endpoint, let $m_r(+v)$ and $m_r(-v)$ be the mapped support-token log-probability margin after positive and negative edits for relation r. We calculate $d_r(v) = \frac{1}{2}[m_r(+v) - m_r(-v)]$ and the relation interaction $I(v) = d_{\text{supported}}(v) - d_{\text{contradicted}}(v)$. The primary contrast subtracts the mean interaction of eight activation-frequency-matched random supports from the selected-support interaction, averaging within source before 10,000 paired source-group bootstrap draws. Both edit signs are applied for every relation and response mapping. A positive pooled interval alone is insufficient: the direction must remain positive under both mappings and cannot be explained by a common same-sign output shift.

A.13 Same-source output-mapping audit

The later mapping audit retains the 42 source groups and the frozen Gemma Top-8 SAE support from Appendix A.12, but adds A/B, X/Y and red/blue to YES/NO. Each codebook is evaluated with both choices of which token means "supported". The model, source records, candidate identities, question polarities and L18 last-prompt-token position are held fixed within a model. The added codebooks were evaluated after the original YES/NO confirmation results were known; their 42-group results are retrospective sensitivity analyses, not new independent confirmation. The selected support is not reselected per codebook. Gemma uses eight activity-matched random SAE supports. Both edit signs are evaluated at 2% of each prompt state's norm. The unedited endpoint is the next-token logit margin of the two code tokens, oriented as supported minus contradicted; its within-source AUROC

Table 7: Complete same-source output-mapping audit. Each cell uses 42 source groups. Columns are within-source AUROCs for the frozen internal readout and unedited two-token output margin, followed by selected-minus-control relation interaction. Values are point estimates; source-bootstrap intervals for the principal contrasts appear in Table 2 and the archived evidence-chain CSV. These are nominal comparisons without multiplicity correction.

Model	Supported code	Internal	Endpoint	ΔI
Gemma	A	0.688	0.643	+0.0072
Gemma	B	0.690	0.640	+0.0060
Gemma	X	0.682	0.524	+0.0011
Gemma	Y	0.652	0.702	+0.0072
Gemma	red	0.649	0.646	+0.0172
Gemma	blue	0.679	0.661	+0.0051
Gemma	YES	0.679	0.702	+0.0482
Gemma	NO	0.696	0.369	-0.0412
Qwen	A	0.571	0.527	-0.0102
Qwen	B	0.625	0.530	-0.0007
Qwen	X	0.595	0.533	+0.0115
Qwen	Y	0.613	0.524	+0.0020
Qwen	red	0.601	0.527	+0.0040
Qwen	blue	0.583	0.542	+0.0051
Qwen	YES	0.595	0.545	+0.0040
Qwen	NO	0.619	0.491	-0.0013

is computed before any patch. Its ability to discriminate the relation is a necessary endpoint check, not a test of free-generation correctness.

For a second-model stress test, Qwen2.5-1.5B-Instruct uses the same 20 development and 42 test source IDs, but its own development-fitted residual ridge direction (fixed penalty 100) and eight coordinate-permutation directions with the same coefficient spectrum. The Qwen development out-of-fold internal AUROC is 0.538 [0.497,0.578]. This weak estimate passed the script's point-estimate gate of AUROC > 0.5, not a lower-bound gate. The SAE and dense support spaces, as well as their control families, cannot be compared as equivalent interventions. The Qwen model outcomes were collected in this extension, but its source split had been examined historically.

For each of the two models, 42 sources $\times$ two candidates $\times$ two question polarities $\times$ four codebooks $\times$ two mappings give 1,344 prompt conditions. On every condition, the selected and eight control directions receive positive and negative patches. The source-level relation interaction is $I(v) = d_{\text{supported}}(v) - d_{\text{contradicted}}(v)$, with $d_r(v) = \frac{1}{2}[m_r(+v) - m_r(-v)]$ and m_r the mapping-oriented output margin. Table 7 reports the selected-minus-mean-control interaction, not the effect of the selected patch alone. Source-group paired bootstrap intervals use 10,000 draws, conditional on the frozen supports and recorded outcomes; they do not correct for evaluating multiple codebooks. A codebook orientation whose baseline endpoint is not reliably above chance cannot support a target-specific intervention interpretation, irrespective of its patch statistic. Even where baseline endpoint discrimination is retained, an interval spanning zero does not establish an effect or prove its absence.

A.14 COMPLETE OBSERVATION AND READOUT TABLES

Table 8: Recomputed micro collision ceilings at temperature 0.7. Each row contains 200 questions and 3,200 trajectories.

Model	T_0	T_1	T_2	T_3
Gemma-2-2B	0.500	0.856	0.925	0.950
Gemma-2-2B-IT	0.500	0.891	0.942	0.961
Qwen2.5-1.5B-IT	0.500	0.788	0.857	0.956
Qwen2.5-7B-IT NF4	0.500	0.829	0.897	0.943

Table 9: Retrospective common-interface audit on 500 HotpotQA questions with four supplied options. Brackets denote 95% question-bootstrap intervals.

Interface	Within-question AUROC	Selection accuracy
Dense residual	0.939 [0.923, 0.953]	0.858 [0.826, 0.888]
PCA, source-selected rank 64	0.920 [0.902, 0.937]	0.834 [0.800, 0.866]
Final-candidate-token NLL	0.870 [0.847, 0.891]	0.748 [0.710, 0.786]
Top-32 SAE	0.780 [0.754, 0.806]	0.570 [0.526, 0.614]
Post-candidate token entropy	0.655 [0.625, 0.684]	0.396 [0.354, 0.438]
Residual norm	0.425 [0.393, 0.457]	0.184 [0.150, 0.220]

A.15 COMPLETE CROSSOVER RESULTS

Table 10: Sign-balanced crossover on 106 historical source contrasts. Values are mapped logit-margin effects with source-bootstrap 95% intervals. Each source averages three directions per family. The reversed mapping fails the baseline endpoint check and is not confirmatory evidence of specificity.

Quantity	Ordinary mapping	Reversed mapping
Baseline relation gap	0.29245 [0.14505, 0.44752]	-0.22700 [-0.32193, -0.13915]
Selected interaction I	-0.00128 [-0.00737, 0.00472]	0.00825 [0.00069, 0.01592]
Random interaction I	-0.00413 [-0.00973, 0.00157]	-0.00049 [-0.00629, 0.00511]
Selected minus random I	0.00285 [-0.00452, 0.01051]	0.00875 [0.00000, 0.01759]
Selected common shift B	0.01902 [0.01592, 0.02221]	0.00884 [0.00496, 0.01268]

A.16 ADDITIONAL SCOPE AND DECISION CONSIDERATIONS

Several analyses reuse inspected corpora; disjoint fitting does not make them prospective. Natural-candidate evaluation uses a shared pool and exact match, while the 17-question relation gate is low-powered and operational. The 106-source crossover is separate. Budgets omit some training, annotation and audit costs; intervals omit analysis history, and inconclusive results are not equivalence tests. In 2Wiki, 550/960 candidates hit the 32-token cap without EOS, confounding length controls. The crossover changes model and readout position, and its reversed mapping fails baseline fidelity; the later 64/42-source codebook split is retrospective. These tests do not establish a natural mechanism or repair policy. An online policy requires decision-time inputs, separate development and cost-matched action baselines; the Qwen readout here is offline.