HYPOTHESIS ANCHORING: EVIDENCE USE WITHOUT BELIEF REVISION ON SELF-GENERATED HYPOTHESES

Anonymous authors
Paper under double-blind review

ABSTRACT

Reasoning models often commit early to an incorrect hypothesis and never revise it, even when their own chain of thought contains the refuting evidence, and even when the correct answer is inserted into their context in plain text. We quantify this failure, which we term *hypothesis anchoring*, with an injection stress test on incorrect GSM8K traces of two open-weight reasoning models, replicated under two annotation protocols for locating the anchored hypothesis. Stating the ground-truth answer inside the model's ongoing reasoning barely increases the rate at which it reaches the correct answer, and a wrong-value injection does not reproduce the gain: insertions act through their content rather than their mere presence, and the model integrates true evidence only weakly. Re-solving the same problems from scratch does markedly better, so the anchored context itself is costly. In roughly a quarter of injected continuations the correct answer appears verbatim in the generated text, yet the model ends its reasoning elsewhere. We then ask where the anchor resides, excluding each candidate with matched controls and reported minimum detectable effects: we find no evidence that anchoring stems from inattention to counter-evidence (it is attended), from evidence unavailability, from any individual attention head or a single representational direction; deleting the hypothesis sentence does no more than a sham deletion, and blocking post-evidence attention to it, with fluency preserved, changes nothing. These nulls are stable across both annotation protocols and both models. A layer-wise attention signature behaves unstably across trace sources and is not advanced as a mechanism. Commitment survives every intervention we test within the anchored context, leaving training-time intervention and scale as the most plausible remaining levers. We release the protocol, annotations under both protocols, and code.

1 INTRODUCTION

Long chain-of-thought reasoning (Wei et al., 2022) is defended on the assumption that it is self-correcting: when an intermediate step fails, later steps can revise it. We measure a failure mode that violates this assumption at its most consequential point: once a reasoning model has generated an early hypothesis, subsequent evidence—including the model's own, and including the ground truth handed to it verbatim—rarely overturns it. We call this *hypothesis anchoring*, kin to the anchoring heuristic of human judgment (Tversky and Kahneman, 1974) but internally generated.

The experiment that frames the paper is simple. Take incorrect traces in which a model is committed to a wrong answer, insert one sentence into the model's own reasoning stating the correct answer, and let it continue deterministically. In our primary protocol the stated answer appears in the continuation in 11.5% of cases against 7.0% under a position-matched neutral control ($p = 7.0 \times 10^{-6}$), while a wrong-value insertion at the same position does nothing ($p = 0.573$): the gain comes from the content of the insertion, not its presence. Re-solving the same problems from scratch reaches 33.4%. The model is not failing to see the answer; it is failing to adopt it.

Prior anchoring results bias the model with an *external* anchor: a numerical hint (Lou and Sun, 2024), a restated prompt (Hao et al., 2026), an edited premise, or a user turn (Oh et al., 2026). Here the anchor is internally generated, and the instrument reverses the usual direction: we ask whether *true*

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.

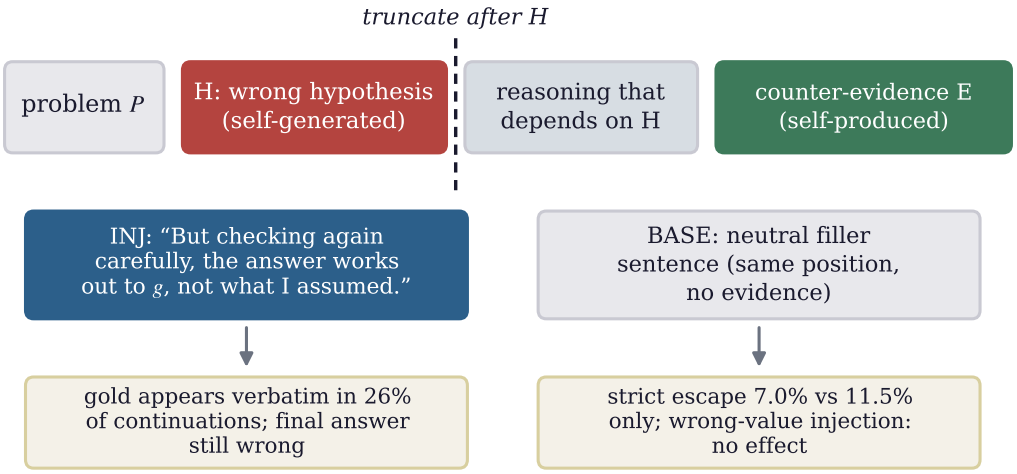


Figure 1: The injection stress test. The model's own trace commits to a self-generated wrong hypothesis (H, red); we truncate after H and insert one sentence stating the ground truth (blue) at a position where a neutral filler serves as control. The continuation rarely adopts the stated answer, which frequently appears verbatim in the generated text while the final answer remains wrong.

evidence injected into the model's own context can overcome an internal commitment—and then locate, by exclusion with matched controls, where that commitment does and does not reside. Our contributions:

- **Instrument and phenomenon.** An injection stress test with content and presence controls, run on-policy on two open-weight models (R1-Distill-Llama-8B (DeepSeek-AI, 2025), Qwen3-8B) and two annotation protocols for H location: evidence acts through content (two-arm $p = 6.9 \times 10^{-6}$ primary; $p = 5.4 \times 10^{-11}$ and $p = 1.4 \times 10^{-8}$ sensitivity), the anchored context costs up to 22 points of solve rate, and in 26.2% of injected continuations the correct answer appears verbatim while the final answer is wrong (§4).
- **Localization by exclusion.** With matched controls and reported MDEs: not perception, not availability, not single heads, not a single direction, not the hypothesis sentence (sham-deletion control, Δ=2.2pp at p=0.040 uncorrected, below our pre-registered 10pp gate), not post-evidence attention to the hypothesis (blockade vs. mass-matched random masking, p=1.0, MDE $\pm$8pp, fluency preserved) (§5).
- **Protocol sensitivity as method.** Two annotation protocols with near-zero inter-annotator agreement on the counter-evidence construct ($\kappa \approx 0.00$ on n=1,805 overlap traces) yield the same qualitative conclusions everywhere, turning annotator relativity from a threat into a robustness result (§6).
- **What survives.** Every inference-time intervention we test inside the anchored context fails; training-time intervention and scale are the remaining levers (§6).

Explicit non-claims. We do not claim to have identified the mechanism of anchoring, and we do not use "integration" as if it named a measured quantity beyond the differential response to true versus wrong insertions. A layer-wise attentional signature that appeared diagnostic in pilot analysis is *not* advanced as a mechanism: it reverses direction between off-policy and on-policy trace sources (§5.1), and it fails significance in the paired pre-registered test on-policy (p=0.24).

2 Related Work

Self-correction and its failures. Intrinsic self-correction without external feedback is unreliable (Huang et al., 2024; Zhang et al., 2025); behavioral drivers are cataloged by Zhang et al. (2025) and Yee et al. (2024), and chain-of-thought may become epiphenomenal past a commitment boundary

(Scalena et al., 2026). These supply the contradicting evidence *externally* and operate at the output level. Our stress test keeps the commitment internal and grades the model on refutations of two provenances—its own, and the ground truth inserted into its own context.

Positioning against prior anchoring and belief-revision work. Table 1 draws the separating axis: in every prior result the anchor or intervention is externally supplied. Here the anchor is internally generated and the instrument tests whether true evidence can overcome it.

Work	Anchor / intervention source	This work
Numerical-anchor studies	external hint	internal (self-generated)
Prompt echoes (Hao et al., 2026)	external, correct, stabilizing	internal, wrong, trapping
Entrenchment under pushback (Oh et al., 2026)	external user turn	no user; self-generated evidence
Premise-edit belief revision	external premise edit	own-context injection
Self-correction failures (Huang et al., 2024; Zhang et al., 2025)	behavioral description	instrument + dissociation + exclusion
Thought anchors (Bogdan et al., 2025)	descriptive saliency	causal probes; signature falsified

Table 1: Every prior anchoring result biases the model with an external anchor; here the model's own hypothesis anchors it against true evidence injected into its own context.

Belief revision. AGM-style normative theory (Alchourrón et al., 1985) and its LLM evaluations (Wilie et al., 2024; Jenkins, 2026; He et al., 2025) measure output-level compliance. We measure an earlier, stronger failure under self-generated commitments. Attention sinks (Xiao et al., 2024) and steering (Zou et al., 2023; Højer et al., 2025) supply our probes; our use of them is deliberately negative-space, with power to bound effects.

3 SETUP

Data. On-policy traces: each model solves GSM8K (Cobbe et al., 2021) (1,319 problems, 8 samples, T=0.6), MATH-500, and AIME-2024; 26,600 traces total (GSM8K accuracy 71.3% for M1). Incorrect traces form the candidate pools.

Annotation protocols. A structured LLM protocol marks, per incorrect trace, the first step committing to a wrong hypothesis the later reasoning depends on (H), whether the trace itself subsequently produces contradicting evidence (C), and whether that evidence is not acted upon (D). Two annotators with near-opposite thresholds on C define our two protocols: *strict* (deepseek-v4-flash; H located in 38.6% of incorrect traces, C in 7.0%, anchored C & D in 6.6%) and *liberal* (Qwen3-8B; C marked on essentially all traces). Their agreement on C and D is $\kappa \approx 0.00$ (n=1,805 overlap traces; exact H-step match 21%, within-one-step 41%). We therefore report all conclusions under both protocols rather than adjudicating a single ground truth.

Pools and conditions. The injection pool (IP) contains incorrect traces with a located H and numeric ground truth; all conditions share truncation immediately after H: BASE (neutral filler), INJ (ground-truth sentence, a fixed first-person verification template), $\text{INJ}_{\text{WRONG}}$ (same template, value $g \pm \delta$, $\delta \in \{1, 2\}$), ER (delete H, restart), SHAM-ER (delete a length-matched random sentence), RESAMPLE (fresh solve from the bare problem). The mechanism pool (AMP) contains anchored (C & D) traces with token spans; its interventions continue after the first counter-evidence. Scoring: strict escape (final answer equals gold) and loose escape; exact McNemar for paired rates; Wilcoxon for ASCE; MDEs reported for every null; primary family Holm-corrected.

4 ACT I: THE PHENOMENON

Three facts organize Table 2. First, **content, not presence**: the direct two-arm comparison INJ vs. $\text{INJ}_{\text{WRONG}}$ gives $+72/-27$, $p = 6.9 \times 10^{-6}$—the same template at the same position helps only

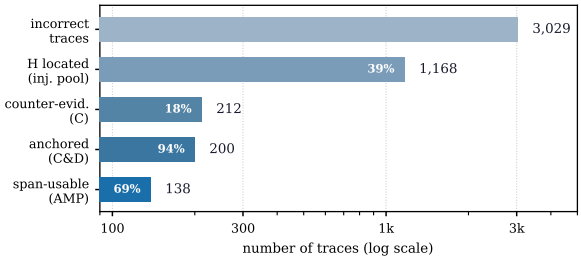


Figure 2: Analysis funnel, strict annotation arm (M1, on-policy GSM8K): 3,029 incorrect GSM8K traces → 1,168 with a located hypothesis step (38.6%; these form the injection pool) → 212 with counter-evidence (C) → 200 anchored (C & D) → 138 mechanism-pool traces after the length guard. The liberal annotator marks C on essentially all traces by construction, so its pools are not funnel-limited.

Condition	Strict escape	vs. BASE (exact McNemar)
BASE (neutral filler)	7.0%	—
INJ (ground truth)	11.5%	$p = 7.0 \times 10^{-6}$
$\text{INJ}_{\text{WRONG}}$ (wrong value)	7.6%	$p = 0.573$
RESAMPLE (fresh solve)	33.4%	vs. INJ $p = 4.6 \times 10^{-41}$
ER (delete H)	9.8%	vs. SHAM +2.2pp, $p = 0.040$
SHAM-ER (delete random)	7.6%	—

Table 2: Primary results: strict annotation arm, M1, on-policy GSM8K, $n = 1{,}168$, greedy continuation.

when its content is true. Second, **the anchored context is costly**: re-solving from scratch reaches 33.4% against 11.5% under injection and 7.0% under neutral continuation; we report RESAMPLE as an upper bound rather than an effect estimate, since the pool is selected on initial failure and fresh solves partially regress toward the problems' conditional solve rates. Third, **seeing is not adopting**: in 17.1% (BASE) and 26.2% (INJ) of continuations the correct answer appears among the numbers the model itself writes (extraction criterion of Appendix B) while the final answer is wrong.

5 Act II: Where the Anchor Does Not Live

Not availability. Ground truth in plain text recovers 11.5% against a 33.4% ceiling (Table 2).

Not the hypothesis sentence. Deleting the annotated H sentence does no more than a sham deletion of a length-matched random sentence: +2.2pp (p=0.040 uncorrected), far below our pre-registered 10pp gate for a localization claim. The commitment is not carried by the identified sentence.

Not post-evidence attention to the hypothesis. During continuation we mask attention from post-evidence positions onto the H span (blockade). Against mass-matched random-span masking and the neutral baseline the escape rates are 11.6% vs. 13.1% and 12.3% (all pairwise p=1.0; n=138; MDE ±8pp), with continuation fluency unchanged (mean NLL 0.24–0.25). Attention onto the hypothesis does not maintain the anchor at this resolution.

Not perception, not heads, not directions. Counter-evidence receives substantial early attention (below), and single-head (Nam et al., 2025) and single-direction probes from our pilot (mid-layer ablation with matched controls; residual steering at fluent doses) produced only nulls with reported MDEs there; both are superseded here by the blockade, which subsumes the attention pathway at the span level.

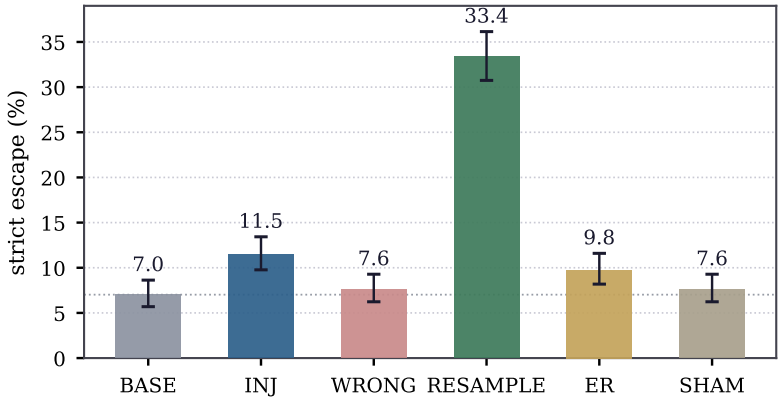


Figure 3: Primary results (strict annotation arm, $n = 1{,}168$): strict escape by condition with Wilson 95% intervals. True-value injection helps; a near-miss wrong value at the identical position does not; discarding the context (RESAMPLE) dominates every in-context condition.

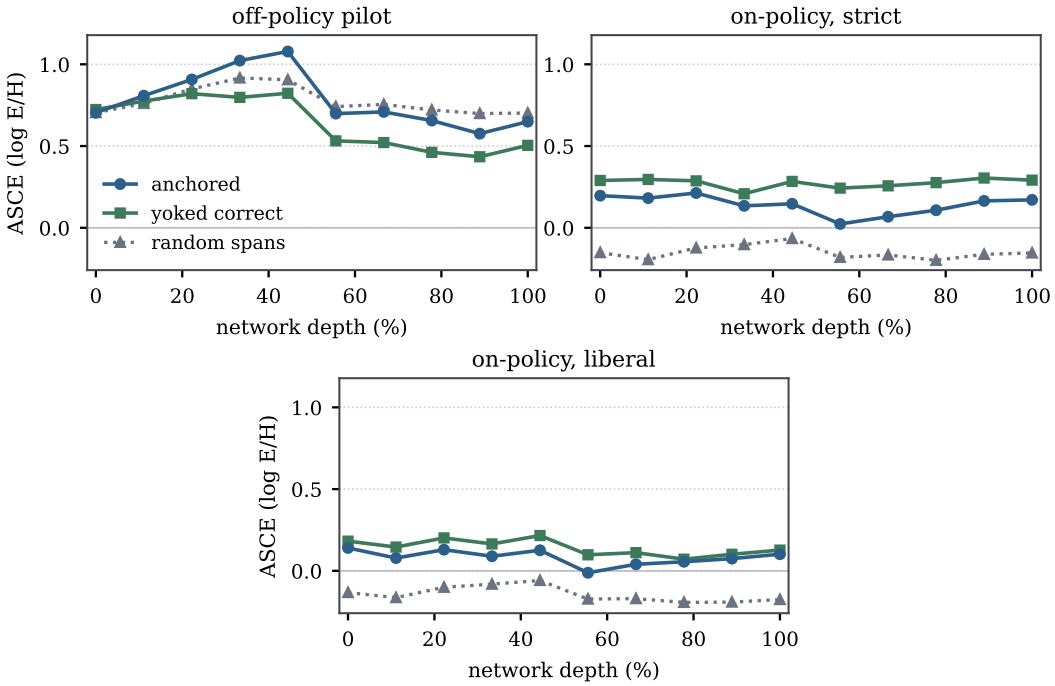


Figure 4: Layer-wise ASCE (log evidence-to-hypothesis attention ratio; positive = attention favors evidence) across the three settings. The claimed anchoring signature—anchored declining while correct does not—appears only in the middle panel and misses the paired significance gate; the random-span control stays flat or negative everywhere. In the liberal panel the yoked-correct series has no per-layer aggregation and is shown as its per-trace early→late means. The quantity is not advanced as a mechanism.

5.1 The attentional signature is not a mechanism

Our pilot (off-policy traces; 66 anchored vs. 38 yoked correct, 30 same-problem pairs) reported a monotone depth decline of the evidence-to-hypothesis attention ratio (ASCE) as the signature of anchoring. Three settings kill this claim as stated. (i) On those off-policy traces, yoked correct traces declined *more* (paired $\Delta = -0.17$, p=0.005): the signature did not discriminate. (ii) On-policy with the strict protocol, anchored traces decline marginally ($+0.041$, p=0.050, n=138) while yoked correct traces do not (-0.019, p=0.87, n=56 unique), and the exact-yoked paired difference misses

significance widely ($\Delta = +0.127$, p=0.24, n=55). (iii) On-policy with the liberal protocol, both anchored and correct traces decline, the correct ones more so (Appendix D). A quantity that flips with trace provenance and annotation protocol, supported in one of three settings and short of significance even there, is not advanced as a mechanism; combined with the blockade null we do not advance attention as the maintaining pathway.

6 Act III: What Survives

Table 3 is the paper's close. Every inference-time intervention *inside* the anchored context fails under both protocols: injecting true evidence helps only weakly, deleting the hypothesis sentence does no more than sham surgery, and blocking attention to it changes nothing. The two levers that do move the rate are discarding the context entirely (RESAMPLE) and, prospectively, training-time intervention—constructing traces that practice contradict, abandon, and re-derive, in the spirit of consider-the-opposite debiasing (Lord et al., 1979) but at the trace level; answer-level ensembling (Wang et al., 2023) does not touch the anchored context. Whether commitment attenuates with scale remains open; the instrument we release makes that a one-line experiment.

Quantity	M1 strict (n=1,168)	M1 liberal (n=3,738)	M2 liberal (n=1,359)
BASE → INJ	7.0→ 11.5 (p=7.0 × 10^{-6})	9.9→ 16.3 (p=10^{-24})	11.3→ 13.8 (p=0.032)
INJ vs. wrong-value	p=6.9 × 10^{-6}	p=5.4 × 10^{-11}	p=1.4 × 10^{-8}
wrong-value vs. BASE	p=0.573 (null)	+2.9pp (p=10^{-6})	−3.9pp (p=10^{-4})
RESAMPLE	33.4%	34.2%	17.5%
ER − SHAM	+2.2pp	+2.1pp	−0.2pp
blockade H vs. rand	p=1.0 (n=138)	p=0.61 (n=111)	—

Table 3: Sensitivity across annotation protocols and models. The wrong-value arm moves across protocols (null, positive, negative) while the content asymmetry (INJ vs. wrong-value) holds everywhere; the toxicity gap and all exclusion nulls are stable.

7 Limitations

(1) The counter-evidence construct is annotator-relative ($\kappa \approx 0.00$); we report under both protocols rather than adjudicating. (2) The strict arm is M1/GSM8K; M2 and MATH-500 ran under the liberal protocol, and cross cells (M2 strict) are unreplaced by design under deadline. (3) H-step location agrees between annotators at 21% exact / 41% within one step; pool composition therefore varies with the locator, which the protocol sensitivity covers. (4) All nulls carry MDEs ($\pm$2.8pp behavioral primary at n=1,168; $\pm$8pp blockade at n=138); effects below these resolutions are not excluded. (5) Single-turn GSM8K-style settings only; multi-turn and agentic anchoring (Oh et al., 2026) are out of scope. (6) Injection used a single template and near-miss wrong values ($g \pm \{1, 2\}$); template variance and far-miss values were probed only in the pilot.

AI use statement

In accordance with the ICLR policy on AI assistance, we disclose the following uses of large language models: research ideation and execution support, including drafting the experimental protocol, the statistical analysis plan, and the experiment scripts, and operating the automated experimental pipeline on rented hardware; retrieval and discovery, including verifying related-work records against arXiv and OpenReview APIs; and drafting and revising sections of this paper, including the abstract. All reported numbers are extracted mechanically from result files produced by the disclosed pipeline and mapped to the paper via a versioned number contract; the authors take full responsibility for the content of this submission. LLMs were not used to prove mathematical claims, and no AI-generated synthetic dataset is reported in this version.

A Annotation Protocol

Each incorrect trace is split into steps (sentence- and newline-based segmentation). The annotator receives the problem, the ground truth, the model's final answer, and the numbered steps (first 50 steps, 250 characters each), and returns a JSON object with: `A`, the index of the first step committing to a wrong assumption that later reasoning depends on (null if the error is an output slip); `B`, an error type; `C`, whether the trace itself subsequently produces content that explicitly contradicts the wrong step (a failed check, a constraint violation, or an explicit recognition of contradiction— hedges and mere restarts do not qualify); `D`, whether C holds and the wrong hypothesis is nevertheless kept or reused; `E`, whether the same wrong approach is revisited; and `F`, a short justification. Decoding is deterministic. The strict annotator (deepseek-v4-flash via API, thinking enabled) yields H on 1,168 of 3,029 incorrect traces (38.6%; these form the injection pool), C on 212 (7.0%), and anchored (C & D) on 200 (6.6%; the mechanism pool samples 139 of these, of which 138 pass the length guard); the liberal annotator (Qwen3-8B served locally, thinking disabled after a quality check showed enabled thinking to be prohibitively slow and disabled thinking to mark C on nearly all traces) defines the opposite operating point. Their agreement on C and D on n=1,805 overlaps is $\kappa \approx 0.00$; exact H-step agreement is 21% (41% within one step). We do not adjudicate: all conclusions are reported under both protocols (Table 3).

B Injection Conditions and Scoring

All conditions truncate immediately after the located H step and append exactly one sentence, then continue greedily for 400 tokens. INJ: "*But checking again carefully, the answer works out to g, not what I assumed.*"; $\text{INJ}_{\text{WRONG}}$: the same sentence with $g \pm \delta$, $\delta \in \{1, 2\}$ drawn per-trace with a fixed seed (near-miss values; far-miss values were probed only in the pilot); BASE: "*Let me continue thinking about this.*"; ER deletes the H step and appends an evidence-restating restart opener; SHAM-ER deletes a length-matched random sentence instead and appends the same opener. RESAMPLE solves the bare problem from scratch. Strict escape requires the continuation's extracted final answer to equal g (boxed/final-answer pattern first, last numeric token second; extraction failures count as non-escape); loose escape requires a valid number different from the original wrong answer. Continuations are screened for degeneration; extraction was audited on 100 random continuations in the pilot.

C Minimum Detectable Effects

For a paired exact McNemar test, the difference detectable at approximately 80% power is $(z_{0.975} + z_{0.80})\sqrt{b+c}/n$ expressed as a proportion, where $b+c$ is the observed discordant count. Observed: INJ vs. BASE 2.8pp; $\text{INJ}_{\text{WRONG}}$ vs. BASE 2.5pp; ER vs. SHAM 2.8pp (all n=1,168); blockade H vs. random-span 7.9pp and vs. BASE 8.4pp (n=138); liberal-arm blockade 9.8pp (n=111). Nulls therefore exclude effects larger than these resolutions.

D Full Sensitivity Tables

Condition	M1 liberal (n=3,738)	M2 liberal (n=1,359)
BASE	9.9%	11.3%
INJ	16.3% (p=7.3×10^{-24})	13.8% (p=0.032)
$\text{INJ}_{\text{WRONG}}$	12.8% (+2.9pp vs. BASE, p=8.5×10^{-7})	7.4% (−3.9pp, p=2.5×10^{-4})
RESAMPLE	34.2% (p=7.3×10^{-88} vs. INJ)	17.5%
ER	11.0%	9.9%
SHAM-ER	8.9%	10.1%

Table 4: Liberal-protocol six-condition tables (on-policy GSM8K). The wrong-value arm is protocol-sensitive in both direction and significance, while the content asymmetry (INJ vs. wrong-value: p=5.4×10^{-11} and p=1.4×10^{-8}) and the toxicity gap hold throughout.

Liberal-arm blockade (n=118/118/111 per condition, e7 pool build): BASE 11.0%, blk-H 11.9%, blk-rand 9.0%; H vs. random-span $+9/-6$, p=0.61 (MDE 9.8pp), fluency preserved. Liberal-arm ASCE: anchored traces (n=293) decline marginally ($+0.109 \rightarrow +0.080$, p=0.044), but yoked correct traces (n=330) decline as well ($+0.172 \rightarrow +0.103$, $p<0.001$)—under this protocol the signature again fails to discriminate.

E Reproduction Details

Models: DeepSeek-R1-Distill-Llama-8B (32 layers) and Qwen3-8B (36 layers), bf16. Generation: vLLM 0.8.5, T=0.6, top-p 0.95, per-sample seeds recorded; GSM8K 8 samples/problem, MATH-500 8, AIME-2024 32 (truncated to 512 traces). Continuations: greedy, 400 tokens (RESAMPLE 1024). Attention analyses: HF transformers 4.51.3, `attn_implementation="eager"`, single forward pass per trace; traces longer than 1,600 tokens are skipped by a mechanical guard (138 of 139 strict-AMP traces pass). The blockade modifies the 4-D causal mask during generation (post-evidence queries barred from H-span keys; random-span control matched to the H span's received attention mass). Statistical tests: exact McNemar (two-sided; `binomtest`), Wilcoxon signed-rank. All scripts, both annotation sets, and a machine-checkable map from every reported number to its result file are released.

F Pre-Registered Decisions

Before the primary runs we fixed: (i) the primary family (INJ vs. BASE; ΔASCE paired decline; blockade H vs. random-span) with Holm correction; (ii) fork gates—an attention-blockade claim required $\geq$10pp over matched masking at corrected $p<0.05$ with fluency preserved (not met: p=1.0); a context-amputation claim required ER $-$ SHAM $\geq$10pp (not met: $+$2.2pp at uncorrected p=0.040); with both gates unmet the paper stands as instrument, quantified phenomenon, and exclusion study, which is the form reported here.

References

C. E. Alchourrón, P. Gärdenfors, and D. Makinson. On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic*, 1985.

P. C. Bogdan, U. Macar, N. Nanda, and A. Conmy. Thought anchors: Which LLM reasoning steps matter? *arXiv:2506.19143*, 2025.

K. Cobbe et al. Training verifiers to solve math word problems. *arXiv:2110.14168*, 2021.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv:2501.12948*, 2025.

Z. Hao et al. Echoes as anchors: Probabilistic costs and attention refocusing in LLM reasoning. *arXiv:2602.06600*, 2026.

Z. He et al. Martingale score: An unsupervised metric for Bayesian rationality in LLM reasoning. *NeurIPS*, 2025.

B. Højer et al. Improving reasoning performance in large language models via representation engineering. *ICLR*, 2025.

J. Huang et al. Large language models cannot self-correct reasoning yet. *arXiv:2310.01798*, 2024.

B. Jenkins. AGM-Bench: Do large language models revise beliefs rationally? *ICLR 2026 Workshop on Logical Reasoning of Large Language Models*, 2026.

C. G. Lord, L. Ross, and M. R. Lepper. Biased assimilation and attitude polarization. *J. Personality and Social Psychology*, 37, 1979.

J. Lou and Y. Sun. Anchoring bias in large language models: An experimental study. *arXiv:2412.06593*, 2024.

J. Oh et al. From belief entrenchment to robust reasoning in LLM agents. *TACL*, 2026.

Q. Zhang et al. Understanding the dark side of LLMs' intrinsic self-correction. *arXiv:2412.14959*, 2025.

D. Scalena et al. Beyond the commitment boundary: Probing epiphenomenal chain-of-thought in large reasoning models. *arXiv:2606.13603*, 2026.

A. Tversky and D. Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157), 1974.

X. Wang et al. Self-consistency improves chain of thought reasoning in language models. *arXiv:2203.11171*, 2023.

J. Wei et al. Chain-of-thought prompting elicits reasoning in language models. *NeurIPS*, 2022.

B. Wilie et al. Belief revision: The adaptability of large language models reasoning. *EMNLP*, 2024.

G. Xiao et al. Efficient streaming language models with attention sinks. *ICLR*, 2024.

E. Yee et al. Faithful and unfaithful error recovery in chain of thought. *COLM*, 2024.

A. Nam et al. Causal head gating: A framework for interpreting roles of attention heads in transformers. *arXiv:2505.13737*, 2025.

A. Zou et al. Representation engineering: A top-down approach to AI transparency. *arXiv:2310.01405*, 2023.