Anchored Turns: A Length-Driven Gradient in Chain-of-Thought Causal Necessity in Sharded Arithmetic Conversations

Dhruv Gupta[*1] **Vasudev Majhi**[*1] **Jagat Sesh Challa**[1]
Murari Mandal[2] **Dhruv Kumar**[1]

[1]BITS Pilani [2]RespAI Lab, KIIT Bhubaneswar

Correspondence: f20221215@pilani.bits-pilani.ac.in

Abstract

Chain-of-thought (CoT) faithfulness is measured almost exclusively in single-turn settings, leaving open how the reasoning-trace-to-answer relationship evolves as conversations lengthen. We introduce **turn-level CoT counterfactual deletion**: at every assistant turn we regenerate the answer at five truncation levels of the current-turn thinking block under greedy decoding, and label a turn *anchored*—a descriptive label for causal redundancy of the trace, not a mechanistic claim—when all parseable regenerations agree on the same answer. Across 67 multi-turn sharded-GSM8K conversations from DeepSeek-R1-Distill-Qwen-7B, the anchored fraction grows 3.9× from short (<10 turns, 7.2%) to long (>20 turns, 28.4%). Across nine reasoning models grouped into five independent training lineages, every lineage anchors more in long conversations than in short ones (sign test 5/5, p=0.031), although one model, Qwen3-32B, is flat. The length gradient and the cross-model analyses are exploratory, not preregistered. A surface-feature predictor flags anchored turns at AUC= 0.71 without additional generation, although its operating point does not transfer reliably from short to long conversations. The evidence is limited to sharded multi-step, derivational reasoning with numeric answers; broader domains remain open. Single-turn CoT benchmarks, by construction, cannot detect this gradient.

1 Introduction

Reasoning models that emit visible chain-of-thought (CoT) traces are deployed in long-running, multi-turn conversational interfaces under the implicit assumption that the trace reflects what the model is computing. Recent work has shown this assumption holds imperfectly in *single-turn* settings: under controlled cue injection, reasoning models acknowledge an influencing hint in their trace between roughly 25% and 60% of the time (Chen et al., 2025; Chua and Evans, 2025); even without adversarial framing, naturally occurring rationalizations appear on 0.4–13% of prompts (Arcuschin et al., 2026). However, it is not clear whether the same picture holds when the model is under sustained conversational pressure: whether the *rate* of causally redundant traces is constant across a conversation, or whether the conversation itself drives the dynamic.

Why this matters for NLP practice. Reasoning models are deployed in long-running multi-turn interfaces — coding assistants, agentic systems, customer-facing chatbots — but are evaluated for faithfulness and safety on benchmarks that are overwhelmingly single-turn. If CoT becomes progressively less causally necessary for the model's answer as conversations lengthen, then single-turn measurements of CoT causal necessity do not describe the long-conversation regime. We do not study those deployment settings directly. We quantify this gap in one controlled setting — sharded multi-step reasoning with numeric answers — show that its direction holds across five training lineages, and provide a cheap online flag that requires no additional generation.

We address this question by extending Lanham-style counterfactual deletion (Lanham et al., 2023) from a one-shot scalar to a turn-level binary time series. At every assistant turn, we regenerate the answer five times under greedy decoding, each time with a different fraction of the current-turn ⟨think⟩ block visible (0%, 25%, 50%, 75%, 100%); all prior conversational context, including prior assistant answers, is held identical across the five regenerations. We label a turn **anchored** when all parseable regenerations agree on the same numeric answer, and *CoT-sensitive* otherwise. **"Anchored" is a descriptive, non-mechanistic label:** it means the current-turn CoT is causally redundant for that

[*] Equal contribution.

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 2536–2558
November 6-10, 2026 ©2026 Association for Computational Linguistics

output under our test, not that the trace is post-hoc rationalisation or that the model is deceptive. We remain agnostic about the underlying mechanism throughout (§5).

Headline finding. The fraction of anchored turns grows roughly 4× from short to long conversations on our primary model (Figure 2). The direction holds at the level of training lineage: grouping nine reasoning models into five lineages, every lineage has a higher anchored rate in long conversations than in short ones (sign test 5/5, one-sided p=0.031; mean +30.5 pp), and six of six do under a finer grouping (p=0.016; §4.3). Per model the evidence is weaker: strict short < medium < long ordering holds in six of nine models and long > short in eight of nine (Table 1). Overall anchoring rates vary from 9.8% to 45.6%, and Qwen3-32B is the only model without a positive late-vs-early contrast. We name this length-driven property the *length-anchoring gradient*, and our predictor and within-conversation analyses (§4) suggest the gradient is a property of the conversation, not just of the model. The length gradient and the cross-model analyses are exploratory. Of our pre-registered tests, H1 and H2 were null, and H2b was significant (partial Spearman ρ=−0.343; §4.5, Appendix C).

Contributions.

1. **Method.** We introduce **turn-level CoT counterfactual deletion** (§3.2): a per-turn binary measurement of CoT causal necessity—to our knowledge the first faithfulness primitive that produces a time series within a multi-turn conversation rather than a single-shot scalar.

2. **Finding.** In an exploratory analysis, the per-turn rate at which CoT is causally redundant rises with conversation length in all five training lineages we test, covering nine reasoning models (lineage sign test p=0.031); in a pre-registered test, conversations with a higher anchored fraction update their answers less often as new evidence arrives, after length control (partial Spearman ρ=−0.343, p=0.007; §4.3, §4.5).

3. **Artifact.** A surface-feature predictor (AUC= 0.71, +0.07 over a length-only baseline; §4.6) flags anchored turns at inference time without additional generation, although its operating point does not transfer reliably from short to long conversations.

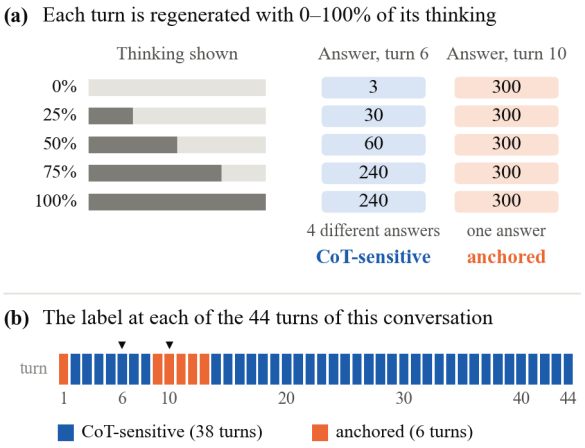


Figure 1: **Turn-level CoT causal necessity test.** (a) Examples of the two turn-level outcomes: **CoT-sensitive** (blue), where different truncation levels of the current-turn CoT produce different answers, and **anchored** (orange), where all parseable truncation levels agree on the same answer. (b) Turn-level classification across a 44-turn sharded-GSM8K conversation, with each turn labeled as CoT-sensitive or anchored.

2 Related Work

Single-turn CoT faithfulness. Lanham et al. (2023) introduced the counterfactual-deletion test we extend here: truncate a ⟨think⟩ block and check whether the answer changes. Follow-up work expanded the picture along several dimensions. Turpin et al. (2023) found that injected biases shift answers without appearing in the trace; Arcuschin et al. (2026) catalogue 0.4–13% rates of naturally occurring rationalization across 15 models; Chen et al. (2025) report that frontier reasoning models acknowledge injected hints in their thinking only 25–39% of the time; Chua and Evans (2025) report that DeepSeek-R1 identifies the causal influence of injected cues 59% of the time, well above the 7% of non-reasoning baselines. At larger scale, Young (2026) test 12 open-weight reasoning models (7B–685B) and find that thinking-token acknowledgment of injected hints (87.5%) systematically exceeds answer-text acknowledgment (28.6%); Shen et al. (2026) introduce a 1,000-trajectory benchmark of instance-level faithfulness across four domains (single-turn, one faithfulness label per prompt). Each of these studies treats each (model, prompt) pair as a unit; none tracks how CoT causal necessity evolves *across turns*.

Multi-turn behavioural failures. Laban et al. (2026) demonstrated that reasoning models lose roughly 39 percentage points of accuracy when

problems are revealed piece-by-piece across turns ("lost in conversation"), with the dominant failure pattern being premature commitment to an early answer. A follow-up by Liu et al. (2026) reframes this as an intent mismatch. Hong et al. (2025) measure how reasoning models reverse their position under user pushback in multi-turn dialogues; Wan et al. (2026) propose mitigation strategies for the inertia these failures produce. None of these works measure the relationship between the reasoning trace and the answer during the multi-turn failure.

Within-generation analogs and the monitorability frame. At the single-generation level, Bogdan et al. (2025) identify sentences within a chain that causally influence the answer ("thought anchors"); Boppana et al. (2026) and Ye et al. (2026) document a within-chain "reasoning theater" / "reasoning horizon" beyond which tokens have negligible causal influence. Our anchored turns are a turn-level analog: the entire current-turn trace can be truncated without changing the answer. Korbak et al. (2025) and Emmons et al. (2025) argue that CoT monitoring is viable only when CoT is computationally necessary for the output; our results quantify how often this condition fails as conversations lengthen. Meek et al. (2025) propose joint faithfulness–verbosity monitoring; our anchoring measure adds a turn-resolved dimension.

3 Method

3.1 Sharded multi-turn conversations

We adopt the sharded-conversation framework of Laban et al. (2026): GSM8K (Cobbe et al., 2021) problems are split into 3–20 shards revealed by a user simulator (Qwen2.5-7B-Instruct; Yang et al., 2024). The simulator re-presents shards across turns rather than advancing strictly one shard per turn, so a conversation runs considerably longer than its shard count: the median is 3.1 assistant turns per distinct shard revealed, and the longest conversation reaches 76 turns. The reasoning model under test — DeepSeek-R1-Distill-Qwen-7B for the primary measurement (Guo et al., 2025) — emits a ⟨think⟩ . . . ⟨/think⟩ block followed by a numeric answer at every turn. This setup creates sustained conversational pressure under which premature commitment is observed empirically (Laban et al., 2026); crucially for our measurement, the entire token context that determines the model's output — the prior assistant answers and the shards revealed so far — is recorded exactly and re-fed verbatim at counterfactual-deletion time, so the only quantity that varies across the five regenerations is how much of the current-turn trace is visible. We collected data in two phases (Phase A: $N = 24$ across four random seeds, max shards 4–10; Phase B: $N = 43$ across three random seeds, max shards 20), giving a combined $N = 67$ conversations and 1,289 raw faithfulness observations.

3.2 Per-turn counterfactual deletion

At every assistant turn t, we extract the thinking block $\mathcal{T}_t$ and the model's answer. We truncate $\mathcal{T}_t$ to five fractions $f \in \{0.00, 0.25, 0.50, 0.75, 1.00\}$ of its original length, close each truncated prefix with ⟨/think⟩, and regenerate the answer greedily (temperature 0, max 128 tokens; raising the cap to 512 tokens changes no turn labels on the tested samples, §4.9). The five regenerations share the same prior context (including all prior assistant answers); the only variable is f. We extract a numeric answer from each regeneration using a shared regex (`(-?[$0-9.,]{2,})|(-?[0-9]+)`) that returns the last matched numeric token after stripping currency symbols and trailing punctuation; if no token is parseable the truncation level is treated as missing. We then define

$$\texttt{is_anchored}(t) = \mathbf{1}[|\{\hat{a}(t,f)\}_{f\in\mathcal{F}}| = 1]\,, \quad (1)$$

where $\hat{a}(t, f)$ is the extracted numeric answer at truncation f and $\mathcal{F}$ is the set of truncation levels yielding a parseable numeric answer. Turns where fewer than three of the five regenerations are parseable are excluded ($\approx 9.5\%$ of turns), giving 1,167 assessable turns from 1,289 raw observations.

Why greedy decoding. The original turns are generated with the DeepSeek-R1 sampling defaults (T=0.6, top-p=0.95). The five counterfactual regenerations use greedy decoding so that *differences across truncation levels* reflect the causal effect of the truncation, not residual sampling noise; with stochastic decoding the five regenerations would diverge even when truncation has no causal effect, spuriously inflating the CoT-sensitive rate. The original answer is recorded but plays no role in the classification: anchoring compares the five greedy regenerations to each other.

3.3 Within-conversation analyses

We address three within-conversation questions — the length gradient (Q1), conversation-level struc-

ture (Q2), and evidence integration (Q3); operationalisations in Appendix A.[1]

3.4 Cross-model evaluation

To test whether the length gradient is specific to one distilled model, we ran the same sharded-GSM8K protocol on eight additional reasoning models with independent random seeds: DeepSeek-R1-Distill-Qwen-14B and DeepSeek-R1-Distill-Llama-8B (Guo et al., 2025; Grattafiori et al., 2024), Qwen3-8B, Qwen3-14B, and Qwen3-32B (Yang et al., 2025), QwQ-32B (4-bit, RL with R1 distillation warmup; Qwen Team, 2025; Yang et al., 2024), GLM-Z1-9B (Team GLM et al., 2024), and Phi-4-mini-reasoning (Xu et al., 2025). Each model runs at a single weight precision throughout (Table 1): 4-bit or 8-bit quantization for seven models, and unquantized 16-bit (bf16) weights for GLM-Z1-9B and Phi-4-mini-reasoning. Because precision is constant within a model, each model's comparison across length bins is made at one precision; precision differences can affect comparisons of overall anchoring level between models. On the primary model, INT4 and INT8 give identical turn labels (§4.9, item (c)). The same canonical `is_anchored` from §3.2 is applied to all model traces, with no model-specific tuning.

4 Results

4.1 A 3.9× length-anchoring gradient

Across the combined $N = 67$ conversations and 1,167 assessable turns, 23.3% of turns are anchored. Conversation length is the strongest single descriptor of this fraction (Figure 2):

- **Short** (<10 turns, $n = 20$): 7.2% anchored
- **Medium** (10–20 turns, $n = 23$): 12.8% anchored
- **Long** (>20 turns, $n = 24$): 28.4% anchored

The short-to-long ratio is **3.9**×; jittered points within each bin show the gradient is not driven by outliers (Figure 2). Conversations collected with longer shard budgets show higher raw anchored rates (27–38% vs. 13%), but this difference disappears after length-stratification, indicating that the gradient follows conversation length rather than collection batch (Appendix B).

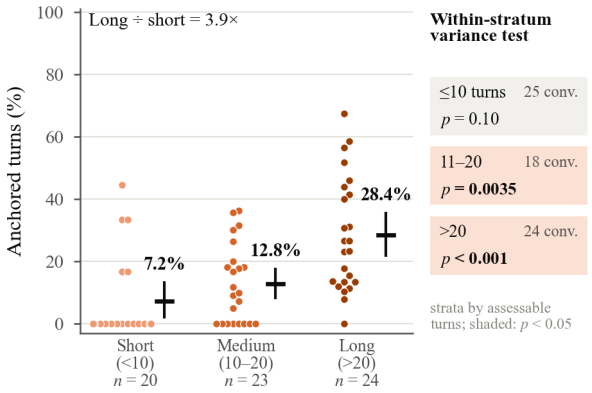


Figure 2: **3.9× length-anchoring gradient** ($N = 67$)**.** Main panel: mean anchored fraction by conversation length ($<$ 10, 10–20, $>$ 20 turns; n = 20, 23, 24). Right panel: variance-test p-values using separate assessable-turn strata ($\leq$ 10, 11–20, $>$ 20; n = 25, 18, 24).

4.2 Testing the information-saturation account

The strongest alternative account of the length gradient is information saturation: as shards accumulate, the answer becomes uniquely determined and the model emits it regardless of CoT. An oracle-onset comparison argues against this account (Appendix F). Anchored and oracle-solvable conversations are disjoint: none of the 64 conversations is both. Of the 42 anchored conversations, none is oracle-solvable inside the 20-shard cap, and 12 anchor at turns 1–3, before significant information has arrived — the reverse of the saturation prediction. Neither this comparison nor the HotpotQA control (Appendix I) rules saturation out as a contributing factor.

4.3 The gradient extends across training lineages

Table 1 reports anchoring rates for the primary model and eight additional reasoning models. Three observations follow from these rows:

Direction. The independent unit is the training lineage, not the individual model, because several rows share a base model or a post-training recipe. We group the nine models by the *Family / training* column of Table 1, which gives six lineages, and obtain five by merging the two R1-distilled rows (Qwen-2.5 and Llama-3 bases) into a single R1-distilled lineage. A lineage effect is the

[1] H1 (run-length persistence), H2 (conversation-level outcome correlation), H2b (length-controlled within-conversation evidence integration), and the conversation ICC together constitute our pre-specified analysis family for the multiplicity correction reported in Appendix C. H2b was framed in our analysis plan as the length-controlled adjunct to H2.

Table 1: **Cross-model anchoring by conversation length.** Short / medium / long = conversations of <10 / 10–20 / >20 total turns; each cell gives the anchored rate and, in parentheses, the number of assessable turns. Quant. = weight precision: 4-bit or 8-bit quantization (4b, 8b), or unquantized bf16 (§3.4). The primary row uses the budget-uniform pool and turn-level rates, so it differs from Figure 2, which uses the full corpus and conversation-level means; both pools are defined in Appendix A. Table 2 (Appendix G) bins the same models by turn position.

Model	Family / training	Quant.	N conv	Short	Medium	Long	S<M<L	L>S	L−S (pp)
R1-Distill-Qwen-7B (primary)	R1 distilled (Qwen-2.5)	8b	51	12.7% (55)	14.4% (215)	30.5% (755)	yes	yes	+17.8
R1-Distill-Qwen-14B	R1 distilled (Qwen-2.5)	4b	18	0.0% (22)	4.7% (64)	30.9% (230)	yes	yes	+30.9
R1-Distill-Llama-8B	R1 distilled (Llama-3)	8b	43	0.0% (53)	3.2% (63)	23.5% (68)	yes	yes	+23.5
Qwen3-8B	RLVR (Qwen3)	8b	49	14.9% (101)	20.5% (88)	45.1% (379)	yes	yes	+30.2
Qwen3-14B	RLVR (Qwen3)	8b	15	15.8% (19)	13.7% (51)	67.7% (99)	no	yes	+51.9
QwQ-32B	QwQ (RL+R1-warmup)	4b	20	0.0% (34)	7.2% (69)	35.5% (186)	yes	yes	+35.5
Qwen3-32B	RLVR (Qwen3)	4b	20	24.4% (41)	15.4% (39)	23.9% (109)	no	no	−0.5
GLM-Z1-9B	RL (GLM-4)	bf16	20	10.8% (37)	10.7% (28)	36.8% (231)	no	yes	+26.0
Phi-4-mini-reasoning	Distilled+RL (Phi-4-mini)	bf16	20	6.1% (33)	16.7% (36)	45.7% (254)	yes	yes	+39.6

unweighted mean of the long-minus-short differences of its member models, taken from the final column of Table 1. The five lineage effects are R1-distilled $+24.1$ pp (three models), RLVR-Qwen3 $+27.2$ pp (three models), QwQ $+35.5$ pp, GLM-4 RL $+26.0$ pp, and Phi-4-mini distilled+RL $+39.6$ pp. All five are positive (sign test 5/5, one-sided $p{=}0.031$); under the six-lineage split, where R1-distilled separates into $+24.4$ pp (Qwen-2.5 base) and $+23.5$ pp (Llama-3 base), the test is 6/6 ($p{=}0.016$). The mean of the five lineage effects is $+30.5$ pp; a one-sample t-test of those five values against zero gives one-sided $p{=}2.7{\times}10^{-4}$ ($\mathrm{df}{=}4$). In Table 1, strict short $<$ medium $<$ long ordering holds for six of nine models and long $>$ short for eight of nine. Two models needed second-seed repair runs to populate the late-turn bin (Appendix G). In the turn-index bins of Table 2 (Appendix G), R1-Distill-Llama-8B is non-monotone (Mid$<$Early but Late $\gg$ both) yet still satisfies Late $>$ Early; Qwen3-32B is the lone model violating Late $>$ Early (flat at $0.94\times$; discussed separately below).

The gradient is not exclusive to distilled models. Qwen3-8B and Qwen3-14B, trained with RLVR (Reinforcement Learning with Verifiable Rewards) rather than by knowledge distillation from R1, also show the gradient. The gradient is therefore not driven by R1-imitation memorisation alone, although the data do not exclude convergent post-training effects across pipelines.

Magnitude varies with architecture and training. At comparable scale, R1-Distill on a Llama-3 base (Grattafiori et al., 2024) anchors $2.4\times$ less than on a Qwen-2.5 base (9.8% vs. 23.4% overall); RLVR-trained Qwen3-14B carries the highest overall rate (45.6%), and within the Qwen3 family the 8B variant anchors at 35.9% on the combined-seed data. Overall anchoring rates span 9.8–45.6%.

The Qwen3-32B flat outlier. Qwen3-32B (4-bit, $N{=}20$, 189 assessable turns) shows no length gradient: Early 20.8% ($N{=}48$), Mid 24.7% ($N{=}85$), Late 19.6% ($N{=}56$); gradient ratio $0.94\times$. The same-size, same-base-family QwQ-32B (RL with R1 distillation warmup) shows a $5.4\times$ gradient on a matched protocol (Table 2). The two 32B rows, matched on scale and base family but differing in post-training pipeline, are consistent with post-training contributing to the effect. **This is a single pairwise contrast; we do not claim post-training is causally isolated**, and we do not generalise beyond this observation: pure RLVR at this scale *can* produce a model without the gradient, whereas RL with R1-style distillation warmup at the same scale preserves it. We treat Qwen3-32B as a boundary condition that constrains the mechanistic interpretation.

Non-reasoning baseline. Qwen2.5-7B-Instruct on the same pipeline gives early-bin anchoring within the reasoning models' range (8.6%, $N{=}35$) but high simulator-side attrition prevented late-bin assessment; full details in Appendix K.

4.4 Conversation identity carries 15% of per-turn variance

A natural alternative to the conversation-level story is that anchoring is per-turn noise with a fixed Bernoulli rate. We reject this account in two complementary ways.

Random-effects ICC. A one-way random-effects ANOVA variance partition across $N{=}66$ conversations (one conversation is excluded for having zero assessable turns) yields $\mathrm{ICC} = 0.152$. This is the closed-form estimate of the random-intercept model $\texttt{anchored} \sim 1 + (1 \mid \mathrm{conv})$ and does not require turn independence. At av-

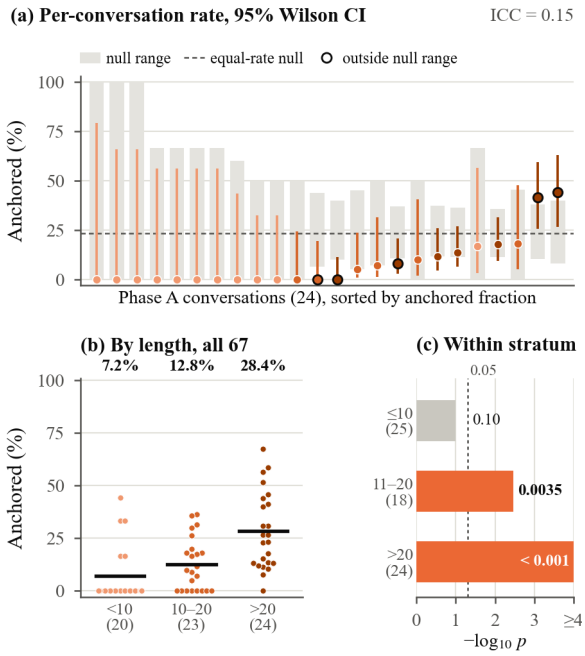


Figure 3: **Conversation-level structure (ICC = 0.15).** (a) Per-conversation anchored fractions with 95% Wilson CIs. (b) Anchored fraction by conversation length. (c) Within-stratum variance-test p-values using assessable-turn strata.

erage cluster size $\bar{n}$=17.7 assessable turns, the design effect is ≈3.54, giving an effective turn-level $N \approx 330$ For nested behavioural data, ICCs in the 0.05–0.20 range are non-negligible cluster effects (Hedges and Hedberg, 2007); ours sits near the top of that range.

Confirmatory variance test. The chi-square variance test against an equal-rate Bernoulli null gives $\chi^2 = 236.9$, $df = 66$, $p \approx 0$. This test assumes turn-independence and is potentially anti-conservative; we read it together with the ICC. Within length strata, the variance test still rejects in the medium ($p = 0.0035$) and long ($p < 0.001$) bins, ruling out the alternative that the structure is solely length-driven (Figure 3).

4.5 Anchoring is associated with a lower answer-update rate

For each conversation we compute the *answer update rate* — the fraction of consecutive-turn pairs where the 0%-truncation answer changes — as a per-conversation signature of failure to incorporate new shards. Conversations with a higher anchored fraction update their context-only answer significantly less often: bivariate Spearman ρ=−0.445, p=4×10^{-4} (N=60); partial Spearman ρ_{partial}=−0.343, p=0.007 after controlling for conversation length. The two measures are computed on disjoint axes (anchoring is a cross-truncation spread within a turn; update rate is a temporal change across turns at a single truncation level), so the association is not a tautology. Full multivariate OLS, Cook's-distance leverage check, the 100%-truncation robustness check, and Figure 11 are in Appendix L.

4.6 A cheap predictor flags anchored turns at AUC= 0.71

A logistic regression on seven surface features extractable from the running conversation log without additional generation achieves pooled GroupKFold ROC-AUC **0.710** (per-fold 0.703 ± 0.033, N=1,162 turns from 63 conversations), versus pooled AUC= 0.643 for a length-only baseline under the same pipeline: a +0.07 AUC lift and a 2.7× drop in per-fold standard deviation. We frame the predictor as a runtime monitoring heuristic, not a mechanistic account (§5); two of the top coefficients are auto-regressive. Trained on conversations of at most 20 turns with its threshold fixed at 90% specificity, and applied unchanged to longer conversations, the predictor does not miss a larger fraction of anchored turns (false-negative rate 73.7% short vs. 67.4% long), but its operating point does not transfer: specificity falls from 90.0% to 81.1%, and both AUC and calibration worsen (Appendix M). Feature list, per-fold breakdown, standardised coefficients, and Figure 12 are in Appendix M.

4.7 Pre-specified null results and what they constrain

Two pre-specified questions did not yield significant findings: H1 (run-length persistence; bootstrap p=0.531) and H2 (conversation-level outcome correlation; r=−0.097, p=0.49). We report them transparently; their failure narrows what the phenomenon *is*, rather than overturning the positive results. Full statistical details, including the H1 bootstrap and the H2 logistic regression, are in Appendix C.

4.8 Specificity to multi-turn presentation

To rule out a generic CoT-redundancy alternative to the conversation-dynamics account, we ran the same five-level test on the matched N=68 GSM8K problems as single-turn prompts. Per-problem anchoring does not transfer between formats (Pearson r=−0.11, p=0.39), so the gradient is not driven

by intrinsic problem properties; full details in Appendix J.

4.9 Robustness controls

Five classes of pipeline confound were each tested directly. (a) *Regeneration token cap:* a regeneration cut off at 128 tokens could yield a different extracted answer, or none, and long conversations could hit the cap more often. Regenerating the same truncation prompts with a 512-token cap, and rebuilding the 128-token condition by cutting each output at token 128, changes zero turn labels across four samples (86 cap-hitting turns and three batches of 160, 141, and 190 turns), so the within-bin anchored rates and the length gradient are unchanged (Appendix E). (b) *Parseability threshold:* re-computing the headline rate under stricter (≥ 5 of 5 parseable) and looser (≥ 2 of 5) exclusion rules moves the rate by only 1.79 pp (23.19–24.98%). (c) *Quantization:* INT4 vs. INT8 runs on six Phase A conversations give 100% per-turn agreement on anchored/CoT-sensitive (184/184; Wilcoxon p=1.0). (d) *Prior-answer ablation:* replacing every prior assistant answer with `[PRIOR ANSWER REDACTED]` across 30 stratified conversations (493 turns per arm) lowers the anchored fraction by 5.24 pp (95% CI $[-9.14, -1.43]$, Wilcoxon p=0.022). Prior answers therefore have a real but limited role: anchoring remains after their removal, and substituting a plausible wrong number for each prior answer does not make the model adopt it ($\leq 4.7\%$ adoption whenever any current-turn CoT is shown; Appendix D). (e) *Phase-B parity check:* 270/270 of Phase B turns reproduce the headline labels under a fixed configuration (shift 0.0 pp; p=1.00), excluding analysis-pipeline drift.

5 Discussion

What we have measured. The five-level test measures one specific property: whether the *current-turn* CoT is causally necessary for the *current-turn* answer, holding all prior context fixed. A turn labelled anchored does not imply that the model is deceptive, that reasoning is irrelevant in general, or that the final answer is wrong: it implies only that the current-turn trace, taken as a whole and across five truncation levels, is insufficient to change that turn's output. We use the descriptive label *anchored* throughout precisely to avoid the stronger and unverified claim that the trace is *post-hoc rationalization*.

Implications for CoT monitorability. Korbak et al. (2025) argue that CoT monitoring is viable as a safety mechanism only when CoT is computationally necessary for the output. Our turn-level measurement operationalises this condition: the fraction of turns where the condition fails grows $\approx 4\times$ from short to long conversations, and the long-over-short contrast is positive in all five training lineages tested (sign test p=0.031). Causal redundancy is not the same as monitor usefulness: a trace that can be deleted without changing the answer may still describe it, and 94.8% of anchored traces state the emitted answer, compared with 92.8% of CoT-sensitive traces. A predictor calibrated on conversations of at most 20 turns does not miss more anchored turns on longer ones (§4.6). The supported claim is that CoT causal necessity changes with length, and that the predictor's operating point does not transfer reliably across this shift.

Candidate mechanisms (hypotheses, not findings). Three mechanisms are consistent with the gradient and our data do not fully discriminate among them: context accumulation (Du et al., 2025), training pressure toward within-conversation consistency (Guo et al., 2025), and information saturation (against which §4.2 is evidence but not a complete discriminator). Prior-answer copying has a real but limited role (§4.9, item (d)): it is not the main explanation for anchoring, and why anchoring increases in longer conversations remains open.

Anchoring is task-modulated. The HotpotQA cross-domain control (Appendix I) shows anchoring at ≈58% with a flat shard-budget gradient, versus 23% with the 3.9× gradient on GSM8K. We interpret two things from this contrast. First, it discriminates against pure information saturation: a saturation account predicts the rate to grow with shard budget on any sharded task with a unique answer, which it does not on HotpotQA. By realized turn index, HotpotQA anchoring is 46.6%, 67.1%, and 53.4% for short, medium, and long turns: non-monotone, and not a reproduction of the sharded-GSM8K gradient. Second, the instrument can be applied outside arithmetic, but the result does not support a general multi-turn dialogue claim: the evidence is limited to sharded multi-step, derivational reasoning with numeric answers, and broader domains, including naturalistic open-ended conversation, code generation, and tool-use trajectories, remain open.

6 Conclusion

Reasoning models are deployed in conversational interfaces under the implicit assumption that their visible chain-of-thought reflects the computation producing the answer. We tested one part of this assumption, whether the current-turn trace is causally necessary for the answer, at the granularity it is actually exercised — turn by turn, inside long conversations — and find that, on sharded multi-step reasoning, the trace is causally redundant more often as conversations lengthen. The measurement was deliberately simple: counterfactual deletion at five truncation levels per turn, yielding a binary time series of when the current-turn trace is causally necessary for the current-turn answer. That simple primitive surfaces a robust effect. The rate at which CoT is causally redundant rises with conversation length, and the direction holds in every training lineage we tested, though not in every model.

The implication for CoT-based monitoring is limited. CoT causal necessity changes with length, but a causally redundant trace can still state the answer, and a predictor calibrated on conversations of at most 20 turns does not miss more anchored turns on longer ones. What does not transfer reliably is its operating point: on long conversations specificity falls, and both AUC and calibration worsen.

What remains open is mechanism. Our data are consistent with multiple plausible drivers and discriminate among none of them cleanly; an oracle-saturation account is constrained by our cross-domain and onset analyses but not eliminated. Closing this gap, and extending the measurement beyond sharded multi-step, derivational reasoning to naturalistic dialogue, code generation, and tool-use trajectories, is where we expect the most useful follow-up.

Data and Code Availability

We release everything needed to reproduce the numbers in this paper: the conversation-collection and user-simulator code, the per-turn counterfactual-deletion harness, the raw regeneration outputs at all five truncation levels for every turn of every model, the derived turn-level label files, and the analysis and figure scripts. **Release:** https://github.com/Dhruv-Git1/Anchored-Turns-A-Length-Driven-Gradient-in-Chain-of-Thought-Sharded-Arithmetic-Conversations.

Limitations

Task and domain. All primary measurements are on sharded GSM8K math problems with a single synthetic user simulator (Qwen2.5-7B-Instruct). Coding, factual question answering, and naturalistic open-ended dialogue likely produce different faithfulness dynamics; the length-anchoring gradient we report is a property of this protocol and should not be projected without test to those settings. Our HotpotQA cross-domain control (Appendix I) shows the absolute anchoring level is task-dependent, and by realized turn index it does not reproduce the GSM8K gradient — findings that strengthen rather than weaken this caveat. Our evidence is limited to sharded multi-step, derivational reasoning and does not support claims about natural dialogue, coding, tool use, or open-ended reasoning.

Models and scale. We test nine open-weight reasoning models of up to 32B parameters. Frontier-scale ($\geq$ 70B) models are not covered. Precision varies across the panel (seven models quantized to 4 or 8 bits, two unquantized in bf16), and our quantization control rules out only the INT4–INT8 precision boundary, not arbitrary precision regimes. A preliminary non-reasoning baseline (Appendix K) matches reasoning models at the conversation onset but does not reach the late-turn regime where the gradient lives; a clean reasoning-vs-non-reasoning comparison remains future work. Our sample is sufficient to detect the gradient and the conversation-level ICC but is underpowered to distinguish a small super-geometric effect from a memoryless null on the run-length test.

Measurement scope. The five-level test labels a turn anchored when the current-turn CoT is causally redundant for that turn's answer. We deliberately do not claim the trace is post-hoc rationalisation or that the model is deceptive: the label measures causal redundancy only. We do not run activation-level mechanistic probes; the candidate mechanisms discussed in §5 are testable hypotheses, not established findings. The primary study is limited to numeric answers in sharded GSM8K: the reported evidence is for numeric multi-step reasoning, not for every answer format or task type. The HotpotQA control uses string and exact-match extraction (Appendix I), but this does not establish performance on arbitrary non-numeric outputs.

Outcome-level claims. Per-turn correctness is ill-defined in sharded GSM8K because early turns

are expected to be wrong before all shards are revealed, and conversation-level correctness is length-confounded (§4.7). We therefore do not make outcome-grounded claims tied to the correctness label. A non-sharded task with well-defined per-turn correctness, or a safety-relevant outcome such as refusal stability under sustained conversational pressure, is the cleanest next step for outcome-grounded analysis. Statistical-multiplicity details for our pre-specified tests are in Appendix C.

Information saturation. The strongest alternative account of the length gradient is that, as shards accumulate, the answer becomes uniquely determined and the model emits it regardless of CoT. Our discriminator experiments (§4.2, Appendix F) — including the cross-domain HotpotQA control and a direct oracle-onset comparison — make a complete saturation account difficult to sustain, but do not rule it out as a contributing factor. A cap-free oracle pass and tasks with explicit determinability annotations are the cleanest follow-ups.

Acknowledgments

The authors would like to thank New Faculty Seed Grant (BITS Pilani), Postman Innovation Labs (BITS Pilani), and Anuradha and Prashanth Palakurthi School of AI, BITS Pilani (APPScAI) for providing Compute and API credits.

We used AI assistants for coding, for preparing figures and tables, and for editing the text; we verified all experiments, analyses and claims, and take full responsibility for the content.

References

Iván Arcuschin, Jett Janiak, Robert Krzyzanowski, Senthooran Rajamanoharan, Neel Nanda, and Arthur Conmy. 2026. Chain-of-thought reasoning in the wild is not always faithful. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 3354–3405. PMLR.

Paul C. Bogdan, Uzay Macar, Arthur Conmy, and Neel Nanda. 2025. Thought anchors: Which LLM reasoning steps matter? *arXiv preprint arXiv:2506.19143*.

Siddharth Boppana, Annabel Ma, Max Loeffler, Raphaël Sarfati, Eric Bigelow, Atticus Geiger, Owen Lewis, and Jack Merullo. 2026. Reasoning theater: Disentangling model beliefs from chain-of-thought. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 9152–9177. PMLR.

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, Vlad Mikulik, Samuel R. Bowman, Jan Leike, Jared Kaplan, and Ethan Perez. 2025. Reasoning models don't always say what they think. *arXiv preprint arXiv:2505.05410*.

James Chua and Owain Evans. 2025. Are deepseek r1 and other reasoning models more faithful? In *ICLR 2025 Workshop on Foundation Models in the Wild*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Yufeng Du, Minyang Tian, Srikanth Ronanki, Subendhu Rongali, Sravan Babu Bodapati, Aram Galstyan, Azton Wells, Roy Schwartz, Eliu A Huerta, and Hao Peng. 2025. Context length alone hurts LLM performance despite perfect retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23281–23298, Suzhou, China. Association for Computational Linguistics.

Scott Emmons, Erik Jenner, David K Elson, Rif A Saurous, Senthooran Rajamanoharan, Heng Chen, Irhum Shafkat, and Rohin Shah. 2025. When chain of thought is necessary, language models struggle to evade monitors. *arXiv preprint arXiv:2507.05246*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z F Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Nature*, 645(8081):633–638.

Larry V Hedges and Eric C Hedberg. 2007. Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, 29(1):60–87.

Jiseung Hong, Grace Byun, Seungone Kim, and Kai Shu. 2025. Measuring sycophancy of language models in multi-turn dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 2239–2259, Suzhou, China. Association for Computational Linguistics.

Tomek Korbak, Mikita Balesni, Elizabeth Barnes, Yoshua Bengio, Joe Benton, Joseph Bloom, Mark Chen, Alan Cooney, Allan Dafoe, Anca Dragan, Scott Emmons, Owain Evans, David Farhi, Ryan Greenblatt, Dan Hendrycks, Marius Hobbhahn, Evan Hubinger, Geoffrey Irving, Erik Jenner, and 22 others. 2025. Chain of thought monitorability: A new and fragile opportunity for ai safety. *arXiv preprint arXiv:2507.11473*.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. 2026. Llms get lost in multi-turn conversation. In *International Conference on Learning Representations*, volume 2026, pages 54738–54778.

Daniël Lakens, Anne M Scheel, and Peder M Isager. 2018. Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, 1(2):259–269.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, and 11 others. 2023. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*.

Geng Liu, Fei Zhu, Rong Feng, Changyi Ma, Shiqi Wang, and Gaofeng Meng. 2026. Intent mismatch causes llms to get lost in multi-turn conversation. *arXiv preprint arXiv:2602.07338*.

Frank J Massey Jr. 1951. The kolmogorov-smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68–78.

Austin Meek, Eitan Sprejer, Iván Arcuschin, Austin J Brockmeier, and Steven Basart. 2025. Measuring chain-of-thought monitorability through faithfulness and verbosity. *arXiv preprint arXiv:2510.27378*.

Qwen Team. 2025. QwQ-32B: Embracing the power of reinforcement learning.

Xu Shen, Song Wang, Zhen Tan, Laura Yao, Xinyu Zhao, Kaidi Xu, Xin Wang, and Tianlong Chen. 2026. Faithcot-bench: Benchmarking instance-level faithfulness of chain-of-thought reasoning. In *International Conference on Learning Representations*, volume 2026, pages 65918–65948.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, and 37 others. 2024. ChatGLM: A family of large language models from GLM-130B to GLM-4 All Tools. *Preprint*, arXiv:2406.12793.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36, pages 74952–74965. Curran Associates, Inc.

Eric-Jan Wagenmakers, Maarten Marsman, Tahira Jamil, Alexander Ly, Josine Verhagen, Jonathon Love, Ravi Selker, Quentin F. Gronau, Martin Šmíra, Sacha Epskamp, Dora Matzke, Jeffrey N. Rouder, and Richard D. Morey. 2018. Bayesian inference for psychology. part i: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, 25(1):35–57.

Yang Wan, Zheng Cao, Zhenhao Zhang, Zhengwen Zeng, Shuheng Shen, Changhua Meng, and Linchao Zhu. 2026. Mitigating conversational inertia in multi-turn agents. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 124404–124431. PMLR.

Haoran Xu, Baolin Peng, Hany Awadalla, Dongdong Chen, Yen-Chun Chen, Mei Gao, Young Jin Kim, Yunsheng Li, Liliang Ren, Yelong Shen, Shuohang Wang, Weijian Xu, Jianfeng Gao, and Weizhu Chen. 2025. Phi-4-Mini-Reasoning: Exploring the limits of small reasoning language models in math. *arXiv preprint arXiv:2504.21233*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380.

Donald Ye, Max Loffgren, Om Kotadia, Linus Wong, and Jonas Rohweder. 2026. Mechanistic evidence for faithfulness decay in chain-of-thought reasoning. In *ICLR 2026 Workshop on Latent & Implicit Thinking – Going Beyond CoT Reasoning*.

Richard J Young. 2026. Lie to me: How faithful is chain-of-thought reasoning in reasoning models? *arXiv preprint arXiv:2603.22582*.

A Experimental details

Infrastructure. All experiments ran on a shared NVIDIA RTX 6000 Ada (49 GB VRAM), Ubuntu 24.04, CUDA 12.4, HuggingFace Transformers 5.5.0, BitsAndBytes 0.49.2. Model weights were cached locally to avoid repeated network downloads.

Hyperparameters. DeepSeek-R1-Distill-Qwen-7B (primary model): temperature 0.6, top-p 0.95, max generation 1500 tokens. Counterfactual regeneration: temperature 0 (greedy), max 128 tokens. Qwen2.5-7B-Instruct (user simulator): temperature 1.0 (simulator), 0.0 (verifier).

Faithfulness token budget. Counterfactual regenerations are capped at 128 output tokens ($\approx$31 s/turn); the metric requires only a parseable numeric value, not a complete explanation. Raising the cap to 512 tokens changes no turn labels on the tested samples (Appendix E).

Data collection. We collected data in two rounds. The first round used shorter shard budgets (maximum 4–10 shards per conversation, four independent random seeds, N=24 conversations); the second used a longer shard budget (maximum 20 shards, three independent random seeds, N=43 conversations). Deduplication across rounds ensures that no GSM8K task appears in more than one round; the combined dataset is treated as a single corpus throughout all analyses.

Analysis pools. Sample sizes differ across analyses because three pools of conversations are used; we name them here once. The *full corpus* is all 67 conversations and 1,289 raw faithfulness observations, of which 1,167 turns are assessable under the $\geq$3-of-5 parseability rule (§3.2). It underlies the

headline rate (§4.1), Figure 2 and the conversation-level analyses; the ICC drops the one conversation with zero assessable turns (N=66, §4.4), and the outcome analyses use the 53 conversations carrying a correctness label (Appendix C). The *budget-uniform pool* is the 51 conversations collected under a single shard budget (1,025 assessable turns); it is used for the primary row of Table 1 and for the train-short, test-long transfer test (Appendix M), so that the length bins are not confounded with collection budget. The *one-run-per-problem pool* keeps a single conversation per GSM8K problem (63 conversations) and is used for the primary row of Table 2 and for the anchoring predictor (Appendix M). Every cross-model row other than the primary model uses that model's own full set of conversations.

Cross-model runs. Each additional model was run with an independent random seed not used in the primary data collection. Two models (R1-Distill-Llama-8B and Qwen3-8B) required a second run with an extended shard budget to populate the late-turn bin; see Appendix G for details. All cross-model runs applied the same canonical `is_anchored` pipeline as the primary model.

Compute. Approximately 68 GPU-hours for the headline experiments ($\sim$20 hours Phase A, $\sim$48 hours Phase B) plus an additional $\sim$55 GPU-hours for the cross-model panel (the two 32B 4-bit runs contribute $\sim$15 GPU-hours beyond the $\leq$14B panel).

Within-conversation analyses (Q1/Q2/Q3). The five-level test labels each turn {*anchored*, *CoT-sensitive*}, giving a binary time series per conversation. We ask three within-conversation questions.

Q1 (length gradient). Does the anchored fraction grow with conversation length? We bin conversations by total turn count (short $<$10, medium 10–20, long $>$20) and report mean anchored fraction per bin with 95% Wilson confidence intervals.

Q2 (conversation-level structure). Is anchoring distributed evenly across conversations, or do some conversations carry far higher rates than per-turn noise alone would produce? We report the ICC of per-turn anchoring (computed via one-way random-effects ANOVA, the closed-form variance partition of a random-intercept model $\texttt{anchored} \sim 1 + (1 \mid \text{conv})$) alongside a χ^2 variance test against the Bernoulli-equal-rate null. The ICC does not assume turn independence within a conversation; the χ^2 test does and is reported as a more sensitive but potentially anti-conservative confirmatory analysis.

Q3 (evidence integration). Is a higher anchored fraction associated with a lower rate of updates to the model's answer as new shards arrive? For each pair of consecutive turns we compute whether the 0%-truncation answer changes; the per-conversation *answer update rate* is the fraction of consecutive-turn pairs where it does. We then correlate the per-conversation anchored fraction with the update rate using Spearman's ρ, controlling for conversation length via partial correlation.

B Stability across data-collection rounds

Conversations collected with longer shard budgets show higher raw anchored rates (27–38%) than those with shorter budgets (13%), because longer budgets draw more conversations into the high-anchoring long bin. Within matched length bins the two collection rounds produce consistent anchored rates (Figure 4), and the ICC is stable across data accumulation.

C Boundary-condition details

Anchored runs are consistent with a memoryless null. We collected 152 anchored runs (mean run length 1.82 turns). A parametric bootstrap test (10,000 simulated datasets) against the best-fit geometric null gives $p = 0.531$; the run-length distribution does not depart from the memoryless null at this sample size. The standard KS test applied naively to discrete data is invalid because the test statistic is mathematically lower-bounded (Massey Jr, 1951); we report the bootstrap-corrected version and treat the result as inconclusive: anchored bursts may either be genuinely memoryless or require $N \geq 200$ conversations to detect a small super-geometric effect.

Conversation-level anchoring is length-confounded for predicting task failure. Across $N = 53$ conversations with correctness labels, the bivariate Pearson correlation between anchored fraction and conversation correctness is $r = -0.097$, $p = 0.49$. Conversation length predicts both anchoring ($r = 0.42$, $p = 0.002$) and failure ($r = -0.37$, $p = 0.006$); a logistic regression on correctness with both predictors gives $\beta_{\text{anchor}} = +1.63$ ($p = 0.48$) and $\beta_{\text{n_turns}} = -0.131$ ($p = 0.008$). Length is the operative predictor of conversation-level failure; anchoring has no

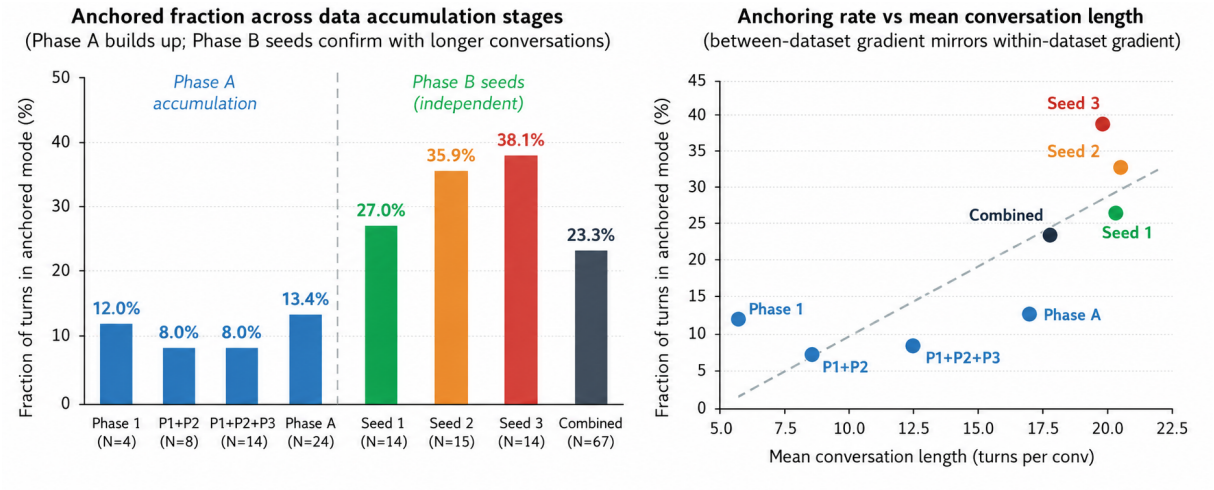


Figure 4: **Anchoring across data accumulation and three independent Phase B seeds.** Left: anchored fraction across all data stages from Phase 1 ($N = 4$) through Phase A ($N = 24$) and three Phase B seeds up to combined $N = 67$. Right: anchored rate vs. mean conversation length per dataset; the between-dataset trend mirrors the within-dataset gradient (Figure 2).

independent association after length control. The H2b length-controlled association test (§4.5) addresses the length-confound directly via partial correlation on the within-conversation update rate, and that test *is* significant after length control — so anchoring carries length-independent signal for the more diagnostic turn-level measure, but not for the conversation-level outcome label.

H1 (run-length persistence) — bootstrap details. Run-length data: 152 anchored runs across the 63 assessable conversations; mean 1.82, observed maximum 10. Parametric bootstrap: fit a geometric distribution to the data ($\hat{p} = 0.55$), draw 10,000 synthetic datasets of 152 geometric($\hat{p}$) samples, compute the KS distance of each to the empirical distribution, and report the upper-tail p-value of the observed KS distance against the simulated null ($p = 0.531$). The standard KS test applied naively to this discrete data would return a guaranteed-significant result because the KS statistic is mathematically lower-bounded by the geometric parameter (Massey Jr, 1951); we therefore do not report it.

H2 (outcome correlation) — Bayesian characterisation. The bivariate Pearson correlation between anchored fraction and conversation correctness ($r = -0.097$, $N = 53$) is accompanied by $\mathrm{BF}_{01} = 4.6$ — moderate Bayesian evidence *for* the null (Wagenmakers et al., 2018) — and a TOST equivalence test (Lakens et al., 2018) against bounds $|\rho| < 0.25$ giving $p_{\text{lower}} = 0.13$, $p_{\text{upper}} = 0.006$. The 95% Fisher-z CI is $[-0.36, +0.18]$. Together these characterise the H2 conversation-level null as genuinely length-confounded rather than merely underpowered; H2b in the main text (partial Spearman after length control on the within-conversation update rate) is the length-controlled adjunct to H2, and it is significant.

Multiplicity correction across pre-specified tests. The Bonferroni-corrected threshold across our four pre-specified tests (H1, H2, H2b, and the conversation ICC) is α_{adj}=0.0125. H2b survives this threshold at both truncation levels (p=0.007 at 0%, p=4×10^{-4} at 100%); H1 (p=0.531) and H2 (p=0.49) do not approach significance, so the correction is immaterial to either. The length gradient (§4.1) and the cross-model analyses (§4.3) were not pre-registered within this four-hypothesis family and are reported as exploratory.

D Mechanism probes and robustness

Answer-token confidence probe (null). For 719 turns from 48 conversations (Phase A plus Phase B seeds 1 and 2), we re-ran the model at 0% and 100% CoT with `output_scores=True` and extracted the margin between the chosen-token logprob and the next-best alternative at the position of the last numeric token in the answer. Anchored and CoT-sensitive turns are statistically indistinguishable on margin at 0% CoT (Mann-Whitney $p = 0.29$, Cohen's $d = 0.05$), at 100% CoT ($p = 0.43$, $d = 0.08$), and on the change in margin between 0% and 100% ($p = 0.72$, $d = 0.03$). Token-

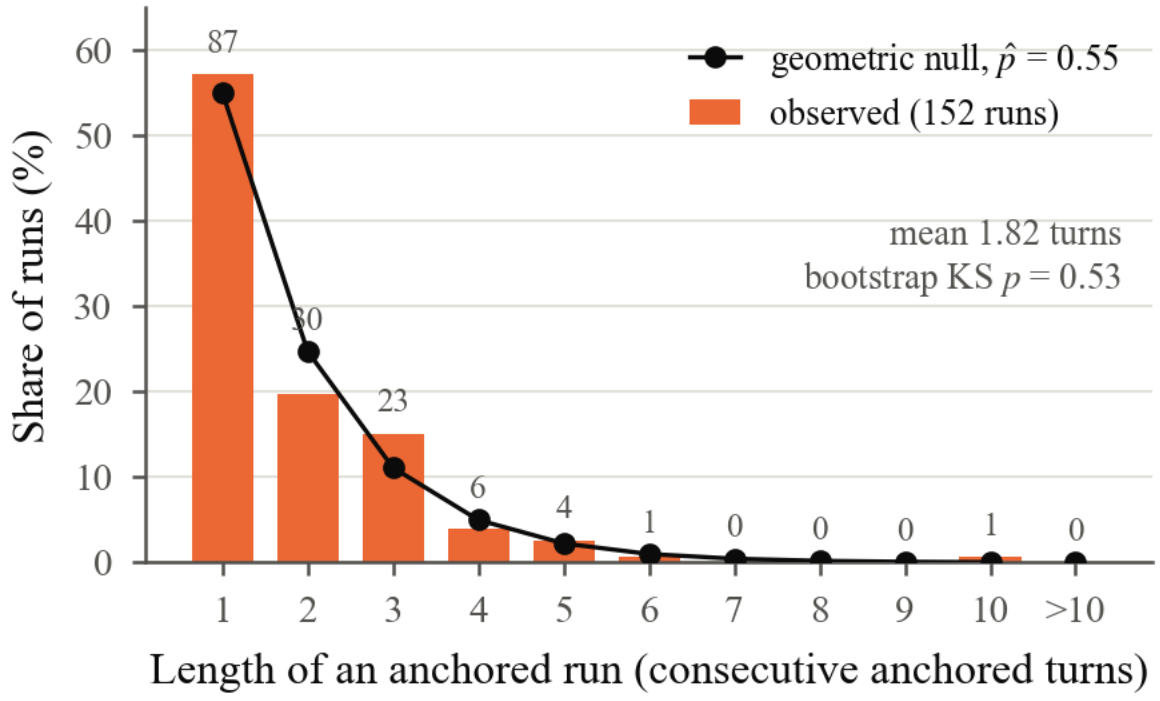


Figure 5: **Anchored run-length distribution vs. geometric null.** Observed anchored-run lengths (bars) compared with the best-fit geometric null ($\hat{p} = 0.55$, black curve). Bootstrap $p = 0.531$; the run-length distribution is consistent with the memoryless null at this sample size.

level confidence is therefore not diagnostic of the anchored/CoT-sensitive distinction (Figure 6); anchoring is an answer-stability phenomenon, not a token-confidence phenomenon.

Commitment-onset probe (null). For each conversation we identified the first anchored turn and a "commitment turn" (first turn at which the model repeats its prior answer on two consecutive turns). Across the 42 conversations with identifiable commitment turns, the Pearson correlation between anchored-onset turn index and commitment-onset turn index is $r = 0.06$, $p = 0.71$. We therefore do not claim that anchored turns are caused by premature commitment; our data do not support the intuitive "early commitment locks in post-hoc reasoning" story.

Thinking-block length is not the driver. The standardised coefficient on log thinking-block length in the predictor (§4.6) is -0.19; conversation length and prior-turn state dominate. At anchored turns in our representative case study the median thinking block is 412 tokens, vs. 387 tokens at CoT-sensitive turns in the same conversation — anchored turns are not anchored because they have nothing to say.

Repetition is a partial mechanism. Of our 272 anchored turns with a readable full-COT answer, 75.8% match the preceding turn's answer, vs. 45.6% of CoT-sensitive turns (ratio 1.66×). Two controls test how much of the phenomenon this accounts for. Redacting every prior assistant answer across 30 stratified conversations (493 turns per arm) lowers the mean anchored fraction by 5.24 pp (95% CI $[-9.14, -1.43]$, Wilcoxon p=0.022): prior answers have a real effect, and anchoring remains after they are removed. Substituting each prior answer with a plausible wrong number tests the copying account directly — if anchoring were copying, the model should adopt the planted number. It does not: planted-number adoption stays at $\leq 4.7\%$ whenever any current-turn CoT is shown, while reproduction of the model's own original answer rises from 22.2% at 0% CoT to 52.6% at full CoT. Prior-answer copying is therefore a minority channel, not the main explanation for anchoring. These controls do not show that the CoT is post-hoc rationalization, and why anchoring increases in longer conversations remains open.

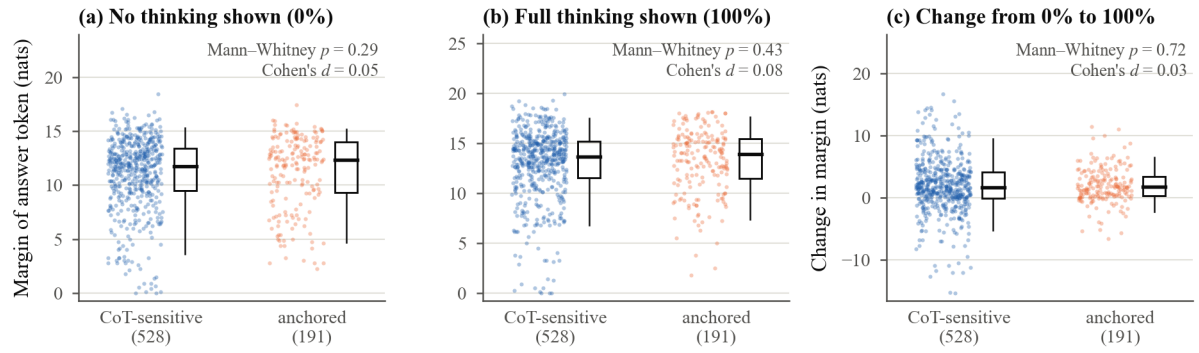


Figure 6: **Answer-token margin distributions are indistinguishable between anchored and CoT-sensitive turns.** (a) Margin at 0% CoT; (b) margin at 100% CoT; (c) change in margin from 0% to 100% CoT.

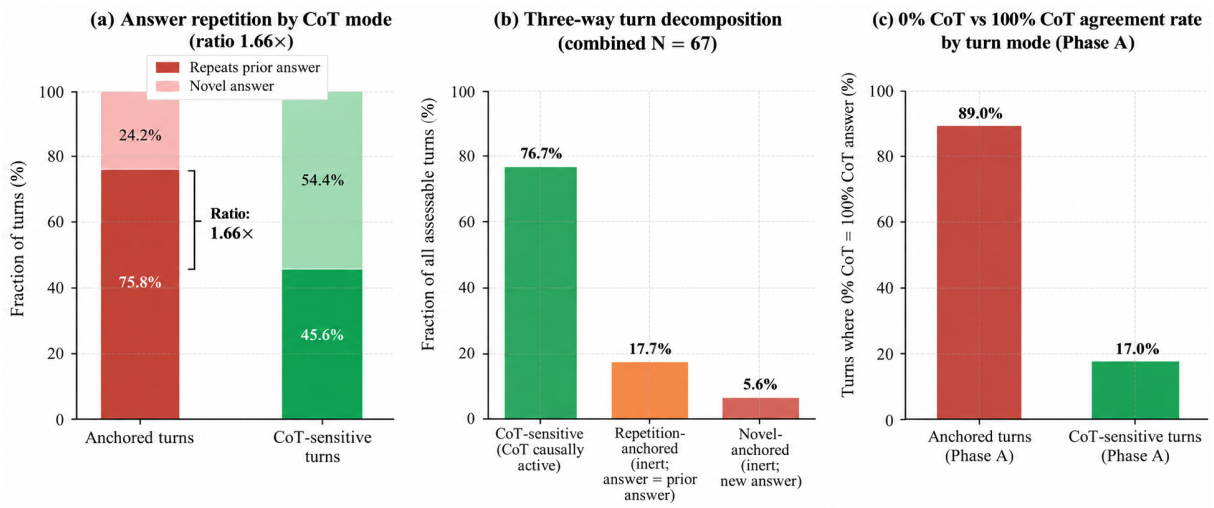


Figure 7: **Anchoring is not reducible to answer inertia.** (a) Prior-answer repetition is more frequent on anchored turns (75.8%) than CoT-sensitive turns (45.6%; ratio 1.66×). (b) Three-way decomposition of turns into CoT-sensitive (76.7%), repetition-anchored (17.2%), and novel-anchored (5.6%). (c) Agreement between the 0%- and 100%-CoT answers is 89% for anchored turns versus 17% for CoT-sensitive turns.

E Regeneration token-budget control

Motivation. Counterfactual regenerations are capped at 128 tokens (§3.2). A regeneration cut off by the cap may end before it states its answer, so the extractor may return an intermediate number or none at all. If long conversations hit the cap more often, the cap could inflate anchoring in the long bin and produce part of the length gradient.

Protocol. For each sampled turn we regenerated every truncation level from the same truncation prompts with greedy decoding and a 512-token cap, holding all other settings fixed. Because greedy decoding is deterministic, the first 128 tokens of a 512-token output are the output the 128-token cap would produce; we therefore rebuilt the 128-token condition by cutting each 512-token output at token 128. We verified this directly: in all 946 regeneration levels where the comparison could be made, the first 128 tokens of the 512-token output matched the corresponding 128-token regeneration exactly. Both versions were scored with the same answer extractor and `is_anchored` rule, and turn labels were compared pairwise.

Results.

Sample	Paired turns	Label changes (128 vs. 512)
Cap-hitting	86	0
Batch A	160	0
Batch B	141	0
Batch C	190	0

No turn changed label in any sample, so on these samples the anchored rate in each length bin, and hence the length gradient, is the same under both caps. The 128-token cap therefore does not inflate anchoring in long conversations and does not pro-

duce the gradient.

F Saturation discriminator

We ran a direct discriminator against the information-saturation account (§4.2). For each of 64 conversations we computed two onset turns: `anchored_onset`, the first turn at which the model becomes anchored, and `solvable_onset_gold`, the first shard prefix k at which an external math oracle (Gemini 2.5 Pro, Comanici et al., 2025, via OpenRouter, 8 samples at temperature 0.7, $\geq$7/8 unanimity matching the gold GSM8K answer) uniquely determines the correct answer. The pre-specified contrast was $\Delta =$ `anchored_onset` $-$ `solvable_onset_gold`: $\Delta < 0$ refutes saturation (model anchors before the puzzle is oracle-solvable); $\Delta \geq 0$ is consistent with saturation.

Result: the two conditions are structurally disjoint. Across the 64 conversations, 42 (65.6%) were anchored and 4 (6.3%) were oracle-solvable to gold within the 20-shard cap; *zero* conversations had both onsets defined (Figure 8A). The 4 oracle-solvable conversations are all short ($\leq$ 8 turns, $\leq$ 6 shards) and never anchored. All 42 anchored conversations have `solvable_onset_gold` $= \emptyset$: even at maximum shard depth, the oracle cannot determine the gold answer from the accumulated shards. A simple delta-distribution test was therefore impossible, but the finding is stronger than the planned probabilistic test: anchoring and oracle-solvability occupy disjoint regions of the dataset.

Anchoring occurs early, before information could saturate. Among the 42 anchored conversations, 12 anchor at turns 1–3 (Figure 8B); 30 of 42 anchor before turn 7. The most extreme cases include conversations that anchor at turn 2 and turn 3 of dialogues lasting 59 and 56 turns respectively. A saturation account requires accumulated information to uniquely determine the answer; at turn 2 the model has seen almost nothing.

Why anchored conversations are oracle-unsolvable. Two structural reasons emerge from inspection of the failure cases. First, 15 of the 64 conversations hit the 20-shard cap; their underlying problems have 21–76 actual shards, and critical quantitative facts (initial amounts, second-day rates) appear only beyond shard 20. The oracle correctly reports `UNDETERMINED` because those facts are genuinely absent from its window. Second, the shards as exposed in the conversation are user-turn messages, not the partitioned sentences of the original GSM8K problem; many conversations repeat one partial fact many times while a baseline quantity needed to solve the problem is never explicitly stated in any shard.

Completeness. Total oracle calls: 4,516; approximately 18% of conversations reached a resource limit before all shard depths were evaluated. The incomplete evaluations all belong to conversations that already showed no oracle-solvable prefix in the completed samples, so the structural disjointness finding is not sensitive to this missing tail. The oracle response parsing failure rate was 14.1%, treated as missing data so as not to inflate the solvable-onset estimate.

G Cross-model late-turn repair runs

Motivation. For R1-Distill-Llama-8B and Qwen3-8B, initial runs produced conversations that terminated early: both models give a confident final answer as soon as the puzzle is locally solvable, after which the user simulator stops issuing shards. This left the late-turn bin (turn $\geq$ 15) with only 17 and 3 assessable observations respectively — insufficient for a credible gradient estimate.

Extended protocol. We ran a second independent set of conversations per model with a larger shard budget (`max_shards`=30, `max_turns`=35). Holding more shards in reserve mechanically lengthens conversations: the simulator distributes them across more turns before the puzzle becomes uniquely determinable, pushing the model past the turn-15 threshold. All other pipeline details — the `is_anchored` definition, the truncation budget, 8-bit quantization, the GSM8K shard pool, and the user-simulator model — were held identical to the first run. Non-overlapping task IDs were enforced across both runs.

Data combination. The two runs are independent samples from the same data-generating process, with the second designed to oversample long conversations. We combined and deduplicated by (task, turn) pair; reported rates reflect the pooled N.

Effect on cross-model estimates. The *direction* of the gradient is unchanged. Late-bin estimates from the initial single runs rested on $N \leq 17$ observations and were substantially inflated by sampling

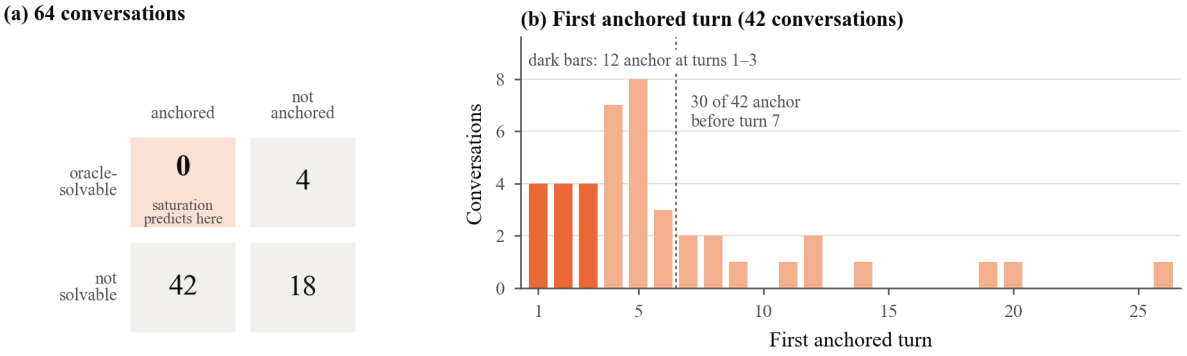


Figure 8: **Saturation discriminator results** ($N = 64$ conversations). **(A)** 2×2 contingency of anchored vs. oracle-solvable status. Under the information-saturation account, anchored conversations should be concentrated in the oracle-solvable cell (top-left); this cell is empty (0/64). **(B)** Distribution of the first anchored turn across the 42 anchored conversations. The dark bars denote turns 1–3; 30/42 conversations anchor before turn 7, including 12/42 at turns 1–3. Together, panels A–B show that anchoring is not explained by the model reaching a uniquely determined answer as information accumulates.

Table 2: **Cross-model anchoring by turn position.** Early / mid / late = turn index < 5 / 5–14 / ≥ 15. The primary row uses the one-run-per-problem pool (63 conversations; Appendix A). Table 1 (§4.3) bins by conversation length and lists each model's training, precision, and conversation count.

Model	Overall	Early	Mid	Late
R1-Distill-Qwen-7B (primary)	23.9%	11.0	23.0	30.9
R1-Distill-Qwen-14B	23.4%	12.1	21.1	31.2
R1-Distill-Llama-8B	9.8%	10.3	5.7	24.1
Qwen3-8B	35.9%	7.7	35.9	54.3
Qwen3-14B	45.6%	8.3	40.8	75.4
QwQ-32B	24.6%	7.0	20.7	37.8
Qwen3-32B	22.2%	20.8	24.7	19.6
GLM-Z1-9B	31.1%	11.5	14.4	49.0
Phi-4-mini-reasoning	38.4%	13.7	36.1	48.4

noise; the pooled estimates are the credible values.

Late-bin counts. Assessable late-bin (turn index $\geq$ 15) observations behind Table 2: primary 457, R1-Distill-Qwen-14B 125, R1-Distill-Llama-8B 54, Qwen3-8B 219, Qwen3-14B 57, QwQ-32B 111, Qwen3-32B 56, GLM-Z1-9B 147, and Phi-4-mini-reasoning 153. Late-bin estimates resting on fewer than 60 observations (R1-Distill-Llama-8B, Qwen3-14B, and Qwen3-32B) are indicative.

H Lineage-clustered cross-model analysis

Correlated model families are not independent replications, so we analyse results by lineage: the independent unit is a training lineage, not an individual model. Conversations are binned by one common length definition: short conversations have fewer than 10 total turns, medium conversations 10 to 20, and long conversations more than 20. Cell counts are assessable turns. The primary-model row uses the budget-uniform pool and reports turn-level cell rates, whereas Figure 2 uses the full corpus and conversation-level means (Appendix A); this is why the primary-model values in Table 1 do not numerically match Figure 2.

How the lineages are formed. Models are grouped by the *Family / training* column of Table 1. That column takes six distinct values, which gives the six-lineage taxonomy. The five-lineage taxonomy merges its two R1-distilled values (Qwen-2.5 base and Llama-3 base) into a single R1-distilled lineage, on the grounds that both come from the same distillation recipe with the same teacher and differ only in base model. A lineage effect is the unweighted mean of the long-minus-short differences of its member models, read from the final column of Table 1; each lineage then counts once, however many models it contains. Table 3 lists the five values, so every statistic in this section can be recomputed from Table 1 alone.

Tests. All five lineage effects are positive (sign test 5/5, one-sided p=0.031), as are all six under the finer taxonomy (p=0.016). The mean of the five is +30.5 pp; a one-sample t-test of those five values against zero gives one-sided $p=2.7\times10^{-4}$ (df=4). Because each lineage contributes exactly one value, the three-model R1-distilled and RLVR-Qwen3 lineages cannot dominate the mean, and the flat Qwen3-32B row is absorbed into its lineage rather than counted as an independent failure to replicate. At the level of individual models, strict

Table 3: **The five lineage effects behind the sign test and the $+30.5$ pp mean.** Each effect is the unweighted mean of its member models' long-minus-short differences in Table 1. Under the six-lineage taxonomy the R1-distilled lineage splits into $+24.4$ pp (Qwen-2.5 base, two models) and $+23.5$ pp (Llama-3 base, one model), and all six remain positive.

Lineage (five-way)	N models	Effect (pp)
R1 distilled	3	$+24.1$
RLVR (Qwen3)	3	$+27.2$
QwQ (RL+R1 warmup)	1	$+35.5$
RL (GLM-4)	1	$+26.0$
Distilled+RL (Phi-4-mini)	1	$+39.6$
Mean of the five lineages		$\mathbf{+30.5}$

short $<$ medium $<$ long monotonicity holds for 6/9, long $>$ short holds for 8/9, and Qwen3-32B is flat. The evidence supports a lineage-level directional result, not a claim that every model follows the same pattern.

I HotpotQA cross-domain control

GSM8K has a particular structure: numeric answers, problems that progressively become uniquely determinable as shards arrive. If anchoring were simply "the answer is now determined and the model emits it," the gradient should appear on any sharded multi-turn task with a unique correct answer. HotpotQA (Yang et al., 2018) fits this description: factual multi-hop QA, one correct answer per question, amenable to the sharded protocol. We therefore ran the same pipeline on HotpotQA as a cross-domain control: if the gradient transfers, information saturation is a plausible mechanism; if it does not, the GSM8K gradient is not a generic "more shards $\Rightarrow$ more determined answer" effect.

Protocol. We applied the `is_anchored` rule of §3.2, with string and exact-match answer extraction instead of the numeric regex, to three independent HotpotQA seeds, each with a different shard-budget cap: `max_turns`$=10$ (short condition, $N=20$ conversations), `max_turns`$=17$ (medium, $N=19$), and `max_turns`$=25$ (long, $N=19$). Same primary model (DeepSeek-R1-Distill-Qwen-7B), same 8-bit quantization, same truncation/regeneration parameters as the GSM8K runs. The only difference per condition is the maximum conversation length the simulator is allowed to run. Code and data are available in the supplementary materials.

Results.

Condition	N conv	Assessable	Anchored	Rate
Short (`max`$=10$)	20	197	113	57.4%
Medium (`max`$=17$)	19	280	175	62.5%
Long (`max`$=25$)	19	350	204	58.3%

Gradient long/short $= 1.02\times$ across shard-budget conditions — strictly flat.

By realized turn index. Binning the same conversations by realized turn index gives anchored rates of 46.6%, 67.1%, and 53.4% for short, medium, and long turns. This pattern is non-monotone and does not reproduce the sharded-GSM8K gradient.

Interpretation. The overall anchoring level on HotpotQA ($\approx 58\%$) is much higher than on GSM8K ($\approx 23\%$). We do not interpret the level: HotpotQA answers are short factoid strings and the model frequently emits the same retrieval from the same prior context regardless of how much of the current-turn CoT is shown. What matters for the saturation argument is the *shape*: a larger shard budget does not produce more anchoring on HotpotQA. A pure information-saturation mechanism predicts a GSM8K-style late-bin rise here; the data show none, flat across shard-budget conditions and non-monotone by realized turn index. We therefore conclude that the GSM8K length-anchoring gradient is not a generic "more shards $\Rightarrow$ more determined answer" effect that any sharded multi-turn task would produce.

Limitations. N is modest per condition (≈ 20 conversations each); we report this as a cross-domain control, not a primary measurement. The high baseline anchoring rate ($\approx 58\%$) on HotpotQA may itself merit study — e.g., whether short-factoid answers are inherently less CoT-dependent than multi-step numeric calculations — but that is outside the scope of this paper.

J Single-turn non-sharded baseline

Summary. To rule out a generic CoT-redundancy alternative to the conversation-dynamics account, we ran the same five-level test on the matched $N=68$ GSM8K problems as single-turn prompts on the primary model. Single-turn anchoring is 11.8% at $T=0.6$ and 5.9% greedy, indistinguishable from the sharded short-conversation rate (binomial $p=0.15$) and significantly below long-sharded ($p=0.0017$). A paired McNemar against any-turn-anchored sharded on the 63 shared task IDs gives

$p=1.4\times10^{-7}$ (39 sharded-only vs. 5 single-only); per-problem anchoring does not transfer between formats (Pearson $r=-0.11$, $p=0.39$), so the gradient is not driven by intrinsic problem properties.

A key alternative to the conversational-dynamics account is that the length-anchoring gradient reflects a property of the problem set: if these GSM8K problems are inherently amenable to CoT-independent answers, the same five-level truncation test should detect a comparable anchoring rate on standard, non-sharded prompts. We tested this on the matched $N=68$ GSM8K problems used in our primary sharded runs.

Protocol. Each of the 68 GSM8K problems is presented to the primary model (DeepSeek-R1-Distill-Qwen-7B, 8-bit) as a single user turn containing the full unsharded problem. The model emits a ⟨think⟩ . . . ⟨/think⟩ block plus an answer; the same five-level truncation test (§3.2) is then applied to that block, with counterfactual regenerations greedy and capped at 128 tokens, as in the sharded runs. The only structural difference between this control and the sharded runs is the format of the original prompt; the `is_anchored` definition is unchanged. Two variants are reported: $T=0.6$ (matching the DeepSeek-R1 sampling defaults used for the sharded original turns) and $T=0$ (fully greedy), to rule out residual sampling noise in the original-turn generation.

Results.

Condition	N problems	Anchored
Single-turn, $T=0.6$	68	11.8% (8/68)
Single-turn, $T=0$ (greedy)	68	5.9% (4/68)
Sharded short (<10 turns)	20 conv	7.2%
Sharded medium (10–20 turns)	23 conv	12.8%
Sharded long (>20 turns)	24 conv	28.4%

At $T=0.6$, single-turn anchoring is statistically indistinguishable from the sharded short-conversation rate (binomial $p=0.15$) and from the sharded medium rate ($p=1.00$). Against the long-conversation rate the difference is significant ($p=0.0017$, binomial). The greedy $T=0$ variant gives an even lower rate (5.9%); the direction is identical, and the cross-temperature paired McNemar is non-significant ($p=0.22$, Cohen $\kappa=0.46$), so sampling temperature on the original turn is not driving the multi-turn gradient.

Paired 2×2 against sharded any-turn-anchored. Restricting to the 63 task IDs present in both conditions:

	sharded anch.	sharded not anch.	total
single-turn anch.	3	5	8
single-turn not anch.	39	16	55
total	42	21	63

The discordant pairs (39 sharded-only vs. 5 single-only) overwhelmingly favour the sharded condition; exact McNemar $p=1.4\times10^{-7}$. Equivalently, of the 42 task IDs that exhibit at least one anchored turn under the sharded protocol, only 3 (7%) also anchor when the same problem is presented single-turn.

Per-problem correlation. We computed each problem's anchored-fraction under the sharded protocol (number of anchored turns over total assessable turns) and correlated it with the binary single-turn anchored indicator. Pearson $r=-0.11$, $p=0.39$; Spearman $\rho=-0.16$, $p=0.23$ ($N=63$). Whether a problem anchors in single-turn does not predict whether the same problem produces anchored turns in the sharded condition; anchoring is therefore not reducible to an intrinsic property of the underlying GSM8K item.

Interpretation. Combining the three observations: (i) the marginal single-turn rate is low and matches the sharded short-conversation rate; (ii) the same problems anchor far more often under sharded presentation, in a paired test that holds task identity fixed; (iii) per-problem anchoring does not transfer between the two formats. The length-anchoring gradient is therefore specific to conversational accumulation, not a generic property of the problem set or of the truncation procedure.

Limitations of this control. Each problem is run once per temperature (single seed); we did not repeat at multiple seeds because the marginal rate of 8/68 (or 4/68 greedy) is unlikely to swing the directional conclusion. The paired comparison uses "any-turn-anchored" on the sharded side, which inflates the sharded rate because longer conversations have more chances to anchor; a more stringent comparison against sharded *turn-1* would tighten the bound on the lower end and a comparison against sharded *last-turn* would tighten the upper end. Both should preserve the directional finding.

K Non-reasoning baseline (Qwen2.5-7B-Instruct)

All nine cross-model rows in Table 2 are reasoning-trained. A natural question is whether the length-

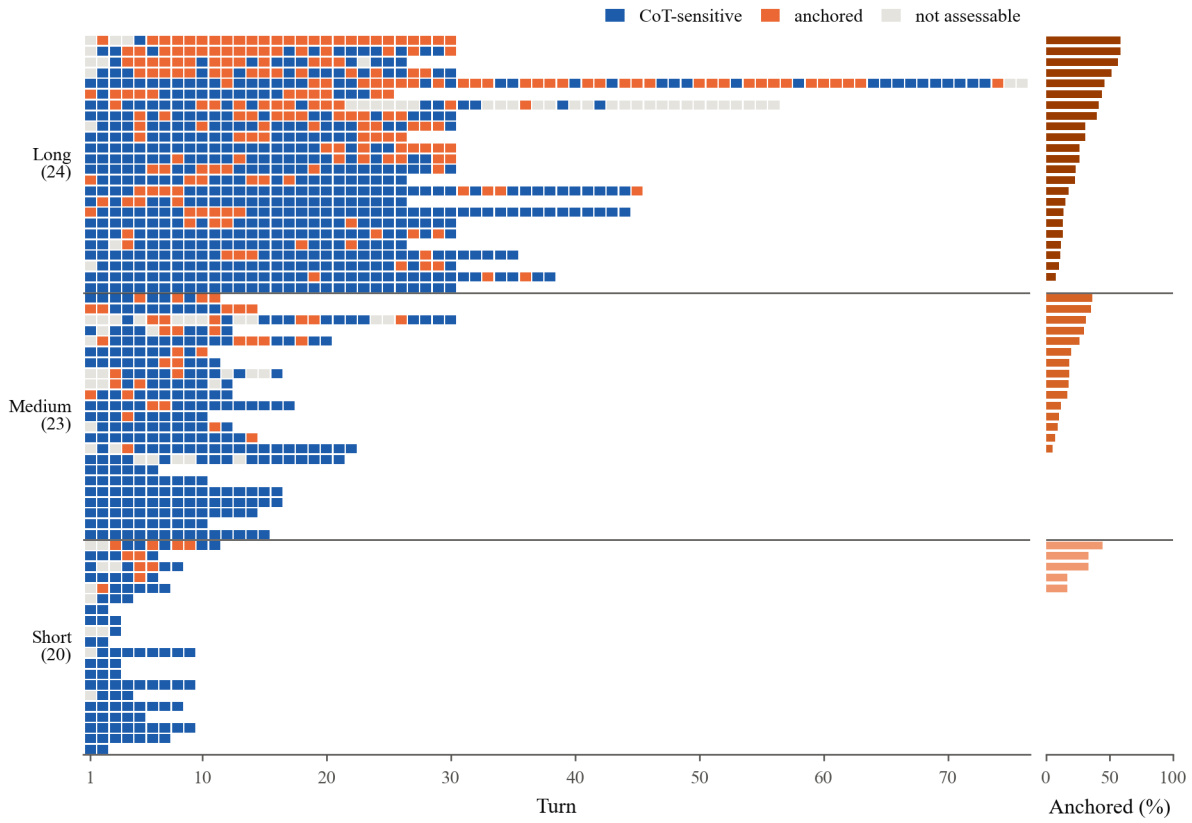


Figure 9: **Per-turn classification across all 67 conversations.** Each row shows one conversation, with turns colored by classification: blue = CoT-sensitive, orange = anchored, and gray = not assessable. Conversations are grouped by length (long, medium, short) and sorted by anchored fraction within each group; the right panel shows the anchored percentage for each conversation. Conversations run longer than their shard count because the simulator re-presents shards (§3).

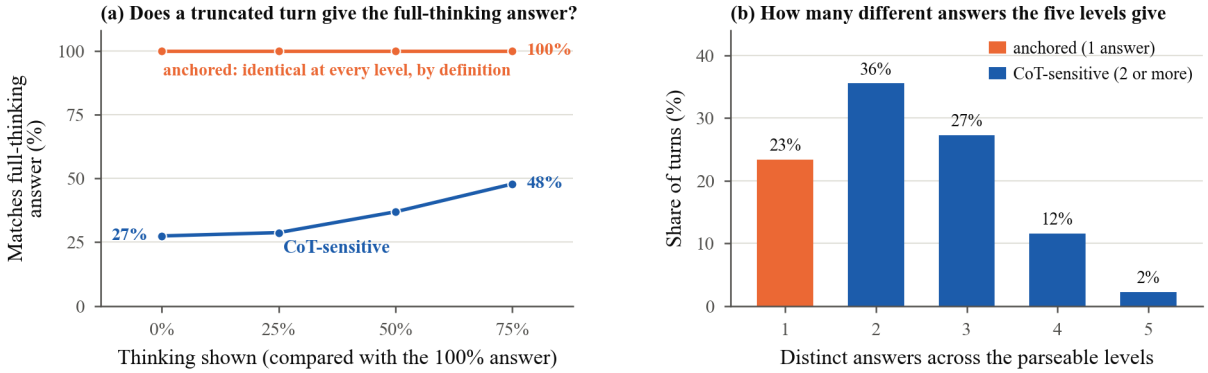


Figure 10: Truncation sensitivity. (a) On CoT-sensitive turns, the share of regenerations that match the full-thinking answer rises from 27% with no thinking shown to 48% at 75%. (b) Anchored turns give a single distinct answer across the parseable levels; CoT-sensitive turns give two or more.

anchoring gradient is a property of reasoning training specifically, or whether any instruction-tuned LLM would show it under the same protocol. Qwen2.5-7B-Instruct is the natural comparison because it is the non-reasoning instruct sibling of the family from which our primary model (R1-Distill-Qwen-7B) was distilled.

Protocol. We applied the same sharded-GSM8K pipeline as the other cross-model rows, using 8-bit quantization and the same truncation and regeneration parameters. Critically, the non-reasoning model was configured to emit a ⟨think⟩ . . . ⟨/think⟩ block before its answer — without this, all five truncation levels are identical and the metric is trivially

100% anchored.

Results. 98 attempted conversations; **22** completed; 76 failed mid-run with a simulator-side parser exception. The simulator could not extract a numeric answer from Qwen2.5's free-form output reliably enough to advance the conversation (the model often produced answers in code blocks, narrative sentences, or refusals that the strict GSM8K verifier rejected). Of the 22 successful conversations, *all* terminated by turn 14 — Qwen2.5-Instruct does not engage in long sharded-puzzle conversations the way reasoning-trained models do.

Numbers.

Bin	N turns	Anchored	Rate
Early (<5)	35	3	8.6%
Mid (5–14)	11	2	18.2% (below N=20)
Late (≥ 15)	0	0	— (no data)
Overall	46	5	10.9%

Interpretation. Early-bin anchoring on the non-reasoning model is 8.6% (N=35 turns), within the early-bin range of the nine reasoning models. This is consistent with the gradient being a conversation-dynamics property rather than a pure reasoning-training artifact: at the conversation onset, before any context has accumulated, the two model classes look the same. The harder question — whether the gradient itself appears in a non-reasoning model — cannot be answered with this data because the gradient regime (turn ≥ 15) was never reached.

Limitations. Attrition. The 76/98 simulator-error attrition is a pipeline infrastructure failure, not a Qwen2.5 refusal; the surviving 22 conversations are not adversarially selected for being short, but they are selected for being parseable by the strict GSM8K verifier. **Verifier.** Yield could be increased by relaxing the verifier so that more Qwen2.5 free-form answers are accepted as numeric. We did not do this because it would break comparability with the reasoning-model runs, which all used the strict verifier. **Conversation-length range.** Reaching the late-turn bin on Qwen2.5 likely requires either (a) fixing the simulator parser, or (b) adopting the repair-run strategy from Appendix G (`max_shards`=30) and accepting whatever yield emerges. A clean reasoning-vs-non-reasoning gradient comparison is the single most informative addition for the "is anchoring a reasoning-model property?" question; we flag it as future work in the Limitations section.

L Evidence-integration analysis details

This appendix expands §4.5. The per-conversation *answer update rate* is the fraction of consecutive-turn pairs where the 0%-truncation answer changes; a low update rate indicates the model's context-only answer is stable despite accumulating shards.

Multivariate OLS. Fitting `answer_update_rate` $\sim$ `anchor_rate` + `n_turns` retains the anchor coefficient as the significant predictor (β=-0.412, p=0.019) while the length term is non-significant after conditioning on anchor rate (p=0.31). The modest attenuation from the bivariate ρ=-0.445 to the partial ρ_{partial}=-0.343 shows that length accounts for only a small fraction of the association.

Leverage check. Removing the two conversations with Cook's distance exceeding $4/N$ leaves the partial Spearman at ρ=-0.441 (p=0.001, N=57), confirming the association is not leverage-driven.

100%-truncation robustness check. Recomputing the update rate using the 100%-truncation answer (full-CoT-included regeneration) instead of the 0%-truncation answer gives bivariate ρ=-0.521 (p=2×10^{-5}, N=60) and partial ρ=-0.442 (p=4×10^{-4}) after length control — *stronger* than the 0%-channel result. The H2b association is therefore not driven by sharing the 0% truncation channel with the anchoring definition.

M Anchoring predictor details

This appendix expands §4.6.

Features. Seven features extractable from the running conversation log without additional generation: turn index, log thinking-block length, prior-answer change, prior-turn anchored status, conversation length, shards revealed, and shard progress.

Protocol. A logistic regression is trained under GroupKFold cross-validation (groups = conversation IDs, preventing within-conversation leakage across folds), N=1,162 turns from 63 conversations. Pooled ROC-AUC is **0.710** (per-fold 0.703 ± 0.033). A length-only baseline under the same pipeline gives pooled AUC= 0.643 (per-fold

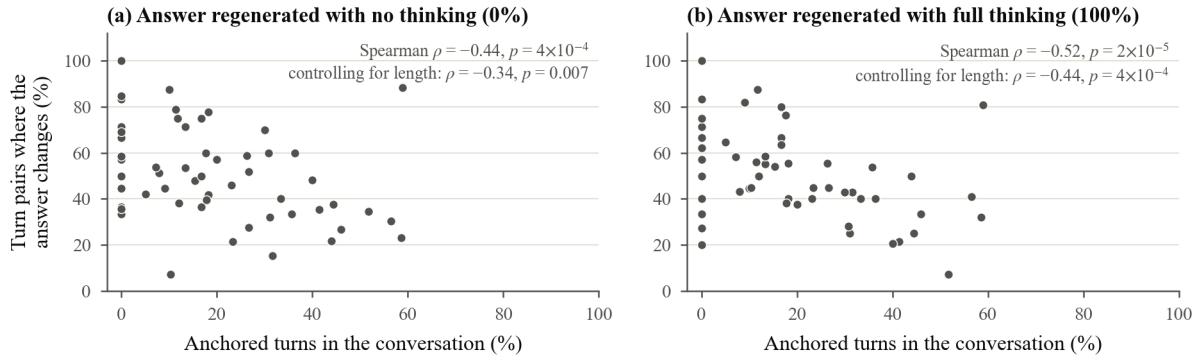


Figure 11: **Anchoring is associated with a lower rate of context-only answer updates.** (a) Anchored-turn fraction versus answer-update rate using the 0%-CoT answer ($N = 60$); Spearman $\rho = -0.444$, $p = 4 \times 10^{-4}$, with partial $\rho = -0.343$, $p = 0.007$ after controlling for conversation length. (b) Same analysis using the 100%-CoT answer; Spearman $\rho = -0.521$, $p = 2 \times 10^{-5}$, with partial $\rho = -0.442$, $p = 4 \times 10^{-4}$ after length control.

0.620 ± 0.088): the six additional features contribute $+0.07$ AUC and a $2.7\times$ drop in per-fold standard deviation.

Top standardised coefficients. Conversation length (β=+0.60), prior-answer change (β=−0.46), and prior-turn anchored status (β=+0.45). The latter two are auto-regressive and propagate serial correlation rather than identify its cause; we therefore frame the predictor as a runtime monitoring heuristic, not a mechanistic account.

Train-short, test-long transfer. We trained the seven-feature predictor on conversations with at most 20 turns, fixed its threshold at 90% specificity, and applied it unchanged to conversations longer than 20 turns.

Metric	Short calibration	Long deployment
Anchored base rate	14.1% (38/270)	30.5% (230/755)
False-negative rate on anchored turns	73.7%	67.4%
Specificity	90.0%	81.1%
AUC	0.659	0.609
ECE	-	0.146

It uses the same primary-model pool as Table 1. The predictor does not miss a larger fraction of anchored turns on long conversations. Its operating point, however, does not transfer: specificity falls from 90.0% to 81.1%, and both AUC and calibration worsen.

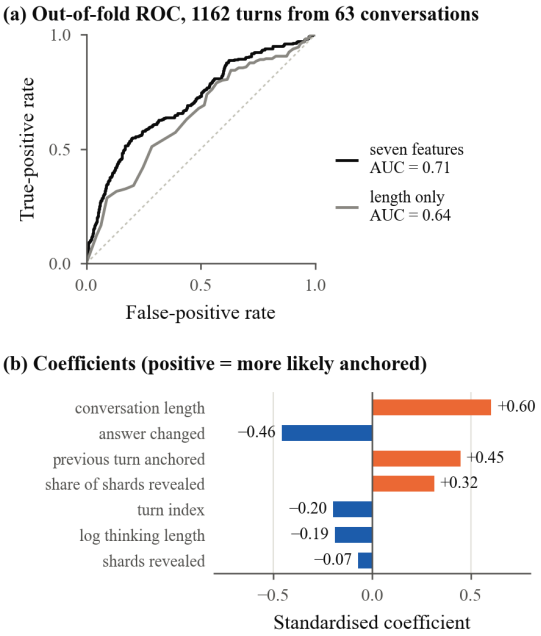


Figure 12: **Predicting anchored turns from 7 surface features.** (a) GroupKFold out-of-fold ROC curves for the seven-feature predictor (pooled AUC= 0.710) and the length-only baseline (AUC= 0.643), using 1,162 turns from 63 conversations. (b) Standardised logistic-regression coefficients; positive values indicate greater likelihood of anchoring.

N Relation to within-generation phenomena

Our anchored turns are phenomenologically similar to within-chain reasoning theater (Boppana et al., 2026) and the reasoning horizon (Ye et al., 2026), but at a different grain: those measure when within-generation tokens stop influencing the final answer; we measure when an entire turn's trace stops influencing that turn's answer. A turn can contain genuine *thought anchors* in the sense of Bogdan et al. (2025) — sentences that drive within-generation computation — while still being anchored in our sense (the current-turn trace as a whole does not constrain the answer). The two concepts are complementary measurements at different grains rather than rival operationalisations.

O Responsible NLP Research Checklist

Scientific artifacts. Datasets used: `sharded_instructions_600.json` from `microsoft/lost-in-conversation`, derived from GSM8K (MIT license). Models used: `deepseek-ai/DeepSeek-R1-Distill-Qwen-7B`, `deepseek-ai/DeepSeek-R1-Distill-Qwen-14B`, `deepseek-ai/DeepSeek-R1-Distill-Llama-8B`, `Qwen/Qwen3-8B`, `Qwen/Qwen3-14B`, `Qwen/Qwen3-32B`, `Qwen/QwQ-32B`, `THUDM/GLM-Z1-9B-0414`, `microsoft/Phi-4-mini-reasoning`, `Qwen/Qwen2.5-7B-Instruct` (user simulator). All artifacts are publicly available on Hugging Face under their respective open-source licenses and are used consistently with their intended research/non-commercial purposes.

Computational experiments. Compute, hyperparameters, and seeds are documented in Appendix A; the total compute envelope is approximately 123 GPU-hours on a shared NVIDIA RTX 6000 Ada. All reported statistics are computed from a fixed analysis script; random-bootstrap procedures use 10,000 replications.

Human annotation and AI assistance. No human-subjects annotation was used; correctness labels come either from the dataset's ground-truth fields or from a deterministic extractor over the model's final answer. AI assistance from a frontier instruction-tuned LLM (used inside an internal AI-coding harness) was used for experimental implementation, code debugging, figure generation, and draft preparation; all decisions of scope, claim, and interpretation were reviewed by the human author.

Ethics and risk. The experiments do not collect human data and do not target any specific deployed system. The safety implications discussed in §5 are about how CoT causal necessity changes with conversation length, we do not release any technique that elicits harmful outputs.