FLAT BY THEOREM: WHAT TRUNCATED CHAIN-OF-THOUGHT ACCURACY CURVES MEASURE

Anonymous authors
Paper under double-blind review

ABSTRACT

A growing literature asks *when* a chain of thought (CoT) decides its answer: truncate it, read out an answer, plot accuracy against the truncation point, read the rise as the moment of commitment. The quantity it measures, the probability that a trajectory ends correct given its prefix, is a bounded martingale under exact continuation, so its expectation is *constant* in the index and equals full-CoT accuracy; the companion agreement curve is a submartingale that rises for every model. Any rise is therefore an artefact; we decompose that of a published-style pipeline, rebuilt on our own traces, into three errors: a readout that intervenes rather than conditions, survivorship, and an index that is not a stopping time. On 8 models and 4 datasets (1 237 682 generations) the exact-continuation curve rises on no arm, equivalent to no change within 0.03–0.05; the forced-readout curve on the *same traces* climbs by +0.03 to +0.64 — by an amount that tracks the ordinary chain-of-thought gain closely enough to predict it: a rule registered on 6 arms before a 14B model was run put its effect at +0.520 [+0.396, +0.643]; it measured +0.633. Such a curve measures what the chain is worth, not when the answer was fixed. The readout dominates: survivorship alone manufactures declines of up to -0.464, but paired within a trace it and index leakage are null. Of 18 chat endpoints audited over one access path, 12 silently ignore assistant prefill, 10 in every sample; and a continuation conditioned on a re-encoded prefix passes every check a single-arm study can run, failing only the one that needs two. What survives is the corrected agreement curve's *shape*, free at half depth between 0.15 and 0.50. Code and records accompany this submission.

1 INTRODUCTION

Chain-of-thought (CoT) prompting lets a language model reason before it answers (Wei et al., 2022), and a natural question follows: at what point in the reasoning is the answer decided? The standard instrument is a truncation curve: sample a CoT, cut it at a sequence of points, elicit an answer from each prefix, and plot how often that answer is correct, or agrees with the model's own final answer, against the truncation point. The curve rises. Lanham et al. (2023) introduced this *early answering* protocol to detect post-hoc reasoning, "reasoning which is generated after the conclusion has already been established"; Ballon et al. (2026) truncate "at fixed token-percentiles" and report that "accuracy and decision commitment consistently increase as the percentage of provided reasoning tokens grows"; Datta et al. (2026) use "forced answer completion, which elicits the model's intermediate predictions at partial reasoning prefixes" to conclude that models *decide early and explain later*; Scalena et al. (2026) locate a "commitment boundary" in the same curve; and Sivakumaran et al. (2026) now train models to improve its area. In every case the curve is read as a statement about the reasoning: how much of the answer is fixed by the time the prefix ends.

This paper is about what that curve measures. The claims made from it are about the trajectory, not about the intervention that reads it: Datta et al. (2026) count the tokens produced "*after the predicted answer has already reached its final value*", Scalena et al. (2026) those generated "*after the answer has already stabilized*". Fix a question and let Y be the answer a sampled trajectory eventually produces. The object such a claim refers to is the conditional law of Y given the prefix, $P_t(a) = \Pr[Y = a \mid c_{1:t}]$, and the natural summary is $K(t) = \mathbb{E}[P_t(a^\star)]$, the probability that a trajectory ends correct given its first t tokens. Our starting observation is elementary and, we will argue, consequential: $(P_t)_t$ is a bounded Doob martingale in the prefix filtration, so $K(t)$ *is constant*

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.

in t and equal to full-CoT accuracy. Conditioning on a prefix of a trajectory that will be completed anyway is, in expectation, uninformative about its success, for a perfect reasoner and a random one alike. The rising curve is therefore a property of the measurement, and our contribution is to say which.

We should be clear about what we are not saying. The forced-answer readout of Lanham et al. (2023) and Datta et al. (2026) is a well-defined quantity — the answer distribution of a model *interrupted* at t and made to answer, a causal effect of an intervention. What it is not is P_t. The two coincide only if forcing an answer at t leaves the answer distribution where continuation would have taken it, and on the same traces at the same indices they do not: a forced readout runs *below* the continuation curve by -0.41 to -0.69 at the shallowest index and climbs back toward it, so the vertical extent it traverses, which is what a published figure displays, is +0.26 to +0.64 accuracy points across the five main arms of Table 1 while the exact-continuation curve traverses nothing. A published curve $K^{\mathrm{force}}(t)$ decomposes exactly as $K^{\mathrm{force}}(t) - \mathrm{Acc} = [K^{\mathrm{force}}(t) - K(t)] + [K(t) - \mathrm{Acc}]$, and the second bracket is zero by theorem: the entire vertical extent of the curve is the intervention. Two further errors compound it. *Survivorship*: including only the traces long enough to reach t plots a curve that must *decline* whenever short traces are more often correct, by an amount we give in closed form and measure at up to -0.464. *Index leakage*: truncating at a fraction of the realised length, or at the last answer switch, uses an index that is not a stopping time, and its bias is a sum of covariances between belief increments and the future the index reveals (both bounded and small on our arms; the readout dominates).

Measuring $K(t)$ correctly requires continuing the exact trajectory under the exact sampling law, which is possible with open weights and completion endpoints but not through most chat interfaces. We audit 18 chat endpoints from four vendors and find the cheapest substitute, a trailing assistant-turn prefill, silently ignored by 12 of them, in every sample by 10. We also derive from the martingale property a family of checks that need no reference arm, each with a positive control realised as an arm on the same base traces: between them they catch a forced readout, a readout that ignores the prefix, a prefix that has been re-encoded, and a base pool built by greedy decoding (Table 18).

Contributions. The martingale structure of conditional answer distributions is known (Falck et al., 2024; Kim et al., 2026; Nikulkov, 2026); $P_t(a^\star)$ is the soft process label of Wang et al. (2023), and the conservation identity we use as an estimator check is Augenblick & Rabin (2021). Our contributions are (i) the identification of the published truncation curves as estimators of $K(t)$ and of $G(t)$, with the nulls they must therefore satisfy: flat for the first, rising for every model for the second; (ii) an exact decomposition of a published curve into readout intervention, survivorship and index leakage, with a closed-form bias for each, and a no-commitment benchmark for last-switch statistics (both Section 2); (iii) a corrected protocol with unbiased estimators and five internal checks, two of which need no reference arm (Section 3); and (iv) a measurement of all three errors on 8 models and 4 datasets, including one model across 4 datasets to separate task from model (Appendix M) and one model held out behind a prediction registered before it was run (Appendix N), plus an audit of 18 chat endpoints (Section 4).

2 The answer-belief process, and what must be flat

Setup. Fix a question with correct answer $a^\star$ in a countable answer set $\mathcal{Y}$ and a model with fixed decoding parameters (temperature, nucleus, cap M). A sampled trajectory is a token sequence $c = (c_1, \ldots, c_T, \mathrm{EOS})$ of random length $T \le M$, padded on the right with the absorbing symbol. A deterministic extractor maps it to a *realised answer* $Y \in \mathcal{Y} \cup \{\perp\}$, where $\perp$ is "no parseable answer" and counts as an ordinary incorrect element. Let $\mathcal{F}_t = \sigma(c_1, \ldots, c_t)$ be the prefix filtration. Two facts are used throughout: $\{T \le t\} \in \mathcal{F}_t$, because an EOS at a position $\le t$ is observable, so T is a stopping time; and Y is $\mathcal{F}_t$-measurable for $t \ge T$. The *answer-belief process* is

$$P_t(a) \;=\; \Pr[\,Y = a \mid \mathcal{F}_t\,], \qquad a \in \mathcal{Y} \cup \{\perp\}, \tag{1}$$

the distribution of the final answer obtained by *continuing the same generation* from the prefix $c_{1:t}$ under the same model and the same decoding law: no prompt change, no forced answer. We study two scalar summaries, the *correctness curve* $K(t) = \mathbb{E}[P_t(a^\star)]$, and the *agreement curve* $G(t) = \mathbb{E}[P_t(Y)]$, which conditions on the trajectory's own eventual answer. Both appear in the literature, usually without distinction; expectations are over trajectories of a fixed question, and averaging over questions changes nothing.

Proposition 1 (Doob martingale). *$(P_t)_{t\geq 0}$ is a bounded martingale with respect to $(\mathcal{F}_t)$, and $P_t = \delta_Y$ for all $t \geq T$.*

Proof. $P_t(a) = \mathbb{E}[\mathbf{1}\{Y = a\} \mid \mathcal{F}_t]$ is the Doob martingale of the bounded variable $\mathbf{1}\{Y = a\}$; the tower property gives $\mathbb{E}[P_{t+1} \mid \mathcal{F}_t] = P_t$, and $\sigma(Y) \subseteq \mathcal{F}_t$ for $t \geq T$ gives absorption. □

This is not new: the same structure is stated by Falck et al. (2024), Kim et al. (2026) and Nikulkov (2026), and $P_t(a^\star)$ is the Monte-Carlo process label of Wang et al. (2023); Lightman et al. (2023) and the value $Q^\pi(s_t)$ of Setlur et al. (2024). The truncation literature has not used the corollary.

Corollary 1 (flat null). *For every deterministic index $t \geq 0$, $K(t) = \mathbb{E}[P_t(a^\star)] = \mathbb{E}[P_0(a^\star)] = \Pr[Y = a^\star] = \mathrm{Acc}$. More generally, $\mathbb{E}[P_\tau(a^\star)] = \mathrm{Acc}$ for every bounded stopping time τ (optional stopping), so adaptive truncation rules that look only at the past—stop at the first sentence boundary, at the first "wait", when an internal confidence probe exceeds a threshold—preserve the null.*

Two readings of Corollary 1 should be kept apart. It does *not* say that reasoning is useless: P_0 already integrates over the whole chain of thought. It says that the fraction of a trajectory one has seen is, in expectation, uninformative about that trajectory's success; the flatness Ton et al. (2024) observed for outcome reward models and read as a defect is, under Corollary 1, the mandatory behaviour of a correctly specified estimator.

Proposition 2 (the agreement curve must rise). *$G(t) = \mathbb{E}\|P_t\|_2^2$, and $\|P_t\|_2^2$ is a submartingale. Hence G is non-decreasing, with $G(0) = \mathbb{E}\|P_0\|_2^2$ (the collision probability of the answer prior) and $G(t) = 1$ for $t \geq M$.*

Proof. $\mathbb{E}[P_t(Y) \mid \mathcal{F}_t] = \sum_a P_t(a) \Pr[Y = a \mid \mathcal{F}_t] = \|P_t\|^2$; conditional Jensen on the convex map $p \mapsto \|p\|^2$ applied to the martingale gives the submartingale property. □

The early-answering metric of Lanham et al. (2023) and the normalised commitment score of Scalena et al. (2026) are *intended* as estimators of G, although both are computed under a forced readout, for which $\|\tilde{P}_t\|^2$ need not be a submartingale and Proposition 2 does not apply; what follows describes the quantity they aim at. Proposition 2 says that this curve rises from $\mathbb{E}\|P_0\|^2$ to 1 for every model and every task, faithful or not, and that its level is set by the concentration of the answer prior. Lanham et al. (2023) report that the 0%-CoT point "is highly predictive of overall early answering … results" and use it as a computational shortcut; Proposition 2 fixes both endpoints of that curve independently of the model, which is consistent with the shortcut, though it does not constrain the shape in between. The increments obey the conservation identity $\mathbb{E}\|P_{t_2}\|^2 - \mathbb{E}\|P_{t_1}\|^2 = \sum_{s=t_1}^{t_2-1} \mathbb{E}\|P_{s+1} - P_s\|^2$ of Augenblick & Rabin (2021); we use it only as an estimator check (Section 3).

2.1 Three ways a measured curve departs from the null

Readout intervention. Published protocols rarely continue the trajectory. They append an instruction such as "*Given the above reasoning, the answer is*" (Lanham et al., 2023; Datta et al., 2026; Ballon et al., 2026) and read the next tokens. Write $\tilde{P}_t$ for the answer distribution under this intervention and $K^{\mathrm{force}}(t) = \mathbb{E}[\tilde{P}_t(a^\star)]$. This is a legitimate causal quantity, the accuracy of a model interrupted at t, but it is not a conditional expectation of anything that will happen if generation proceeds, and nothing forces $\mathbb{E}[\tilde{P}_{t+1} \mid \mathcal{F}_t] = \tilde{P}_t$. The identity

$$K^{\mathrm{force}}(t) - \mathrm{Acc} \;=\; \underbrace{\big[K^{\mathrm{force}}(t) - K(t)\big]}_{\text{intervention effect}} \;+\; \underbrace{\big[K(t) - \mathrm{Acc}\big]}_{=0 \text{ by Corollary 1}} \tag{2}$$

is an identity, but its content is that the second bracket is zero by Corollary 1: the vertical extent of a forced-readout accuracy curve is entirely the effect of the intervention. Both terms are estimable on the same traces, and the first is what we measure in Section 4.

Survivorship. Because T is a stopping time, conditioning on "the trace is still generating at t" is legitimate, but it defines a different estimand at every t, over a shrinking population.

Proposition 3 (survivor curve). *Let* $\pi_t = \Pr[T > t]$, $S(t) = \mathbb{E}[P_t(a^\star) \mid T > t]$ *and* $F(t) = \mathbb{E}[\mathbf{1}\{Y = a^\star\} \mid T \le t]$. *Then* $\mathrm{Acc} = \pi_t S(t) + (1 - \pi_t)F(t)$ *for every* t, *so* $S(t) = \big(\mathrm{Acc} - (1 - \pi_t)F(t)\big)/\pi_t$.

The survivor curve is a deterministic function of the flat null, the survival function and the accuracy of the finished traces; whenever finishers are more often correct than average, $S(t)$ must decline. A pipeline that truncates at an absolute t and skips traces shorter than t plots $S(t)$; that is what happens by default, since a finished trace has no token at position t to truncate at, and it is avoided only by adopting the *absorption convention* explicitly: a trace shorter than t contributes δ_Y at t.

Index leakage. Several protocols truncate not at a fixed t but at an index that depends on the completed trace: a fixed fraction f of its realised length (Lanham et al., 2023; Ballon et al., 2026; Nikulkov, 2026), or the last point at which the elicited answer changed (Datta et al., 2026; Zhang et al., 2026). Such an index ρ is not a stopping time, and optional stopping no longer protects the null.

Proposition 4 (index leakage). *For any random index* ρ *with* $\rho \le M$, *writing* $\Delta_s = P_{s+1} - P_s$,

$$\mathbb{E}[P_\rho(a^\star)] - \mathrm{Acc} \;=\; \sum_{s \ge 0} \mathrm{Cov}\big(\Delta_s(a^\star),\, \mathbf{1}\{\rho > s\}\big). \tag{3}$$

If ρ *is a stopping time every term vanishes; otherwise the bias is the covariance between belief updates and the future the index reveals.*

Proof. Pathwise, $P_\rho - P_0 = \sum_{s \ge 0} \Delta_s \mathbf{1}\{\rho > s\}$. Take expectations and use $\mathbb{E}[\Delta_s] = 0$ from Proposition 1, so $\mathbb{E}[\Delta_s \mathbf{1}\{\rho > s\}] = \mathrm{Cov}(\Delta_s, \mathbf{1}\{\rho > s\})$. When ρ is a stopping time, $\mathbf{1}\{\rho > s\}$ is $\mathcal{F}_s$-measurable and $\mathbb{E}[\Delta_s \mathbf{1}\{\rho > s\}] = \mathbb{E}[\mathbf{1}\{\rho > s\}\mathbb{E}[\Delta_s \mid \mathcal{F}_s]] = 0$. □

Corollary 2 (fraction of realised length). *For* $\rho_f = \lceil fT \rceil$ *with* $f \in (0, 1)$, $\mathbf{1}\{\rho_f > s\} = \mathbf{1}\{T > s/f\}$, *which is* $\mathcal{F}_s$*-measurable for no* s, *and* $\mathbb{E}[P_{\rho_f}(a^\star)] - \mathrm{Acc} = \sum_s \mathrm{Cov}(\Delta_s(a^\star), \mathbf{1}\{T > s/f\})$*: the bias of a length-normalised curve is the covariance between belief increments and* future trace length. *Its sign follows the within-question correlation between length and correctness. It vanishes at* $f = 1$ *for the exact index* $\rho_1 = T$, *where every trace is absorbed at its own answer; on a finite grid the read index is the nearest sampled point, which for* f *near* 1 *is strictly below* T *on a coarse grid and need not vanish.*

Corollary 3 (last-switch statistics and the arcsine law). *Let* $t^\star = \max\{s : A_s \ne A_{s+1}\}$ *be the last index at which an elicited answer* A_s *changes. Model the elicited answer under no early commitment as the sign of a driftless symmetric walk—an assumption, not a consequence of Proposition 1, whose increments are neither symmetric nor identically distributed. Then by Lévy's arcsine law (Lévy, 1940)* $t^\star/T$ *converges to the arcsine distribution on* $[0, 1]$, *with mean* $\frac{1}{2}$ *and about a quarter of its mass below* 0.15*: a process with no early commitment reports, on average, "half of the trace after the decision". Appendix I replaces this idealisation with a simulated belief martingale read through each arm's own grid, which is the null we actually test against.*

Published post-decision fractions of 35–70% (Datta et al., 2026; Scalena et al., 2026) are of this magnitude, and are reported on both sides of it as evidence of the same early commitment.

2.2 A SPECIFICATION TEST THAT NEEDS NO REFERENCE ARM

Let Π be any protocol producing pseudo-beliefs $\tilde{P}_t$: forced readout, a probe, a chat "continue this", a verbalised confidence. If Π measures the exact continuation law at an adapted index then $\mathbb{E}[\Delta\tilde{P}_t \mid \mathcal{F}_t] = 0$, and hence, for any bounded $\mathcal{F}_t$-measurable h,

$$\mathbb{E}\big[\,(\tilde{P}_{t+1}(a^\star) - \tilde{P}_t(a^\star)) \cdot h(c_{1:t})\,\big] = 0. \tag{4}$$

Testing Eq. (4) over a family of h is a specification test computed *inside a single arm*; the simplest member, $h \equiv 1$, tests whether the per-interval increments have zero mean. The condition is Proposition 1 of Falck et al. (2024) and the martingale score of He et al. (2025); both indict the model's rationality, here it indicts the measurement. It needs no reference arm, so a study that has only a forced-readout curve can still test whether that curve is a conditional expectation. Its power, though, is confined to readouts whose increments have non-zero conditional mean, an intervention such as forced readout. It has none against a readout that ignores the prefix, whose pseudo-beliefs are constant in expectation and satisfy Eq. (4) trivially. A second check covers exactly that case.

Proposition 5 (terminal agreement)**.** *Under exact continuation $P_t(a) \to \mathbf{1}\{Y = a\}$ a.s. as $t \to T$ (Lévy's upward theorem, Y being $\mathcal{F}_T$-measurable), so $\mathbb{E}[P_t(Y)] \to 1$ along the trace; under a readout that ignores the prefix, $\tilde{P}_t$ is a draw from P_0 for every t and $\mathbb{E}[\tilde{P}_t(Y)] = \mathbb{E}\|P_0\|^2$ at every depth. The within-trace paired rise between a deep and a shallow index therefore tends to $1 - \mathbb{E}\|P_0\|^2 > 0$ in the first case and is exactly zero in the second. We use only that direction: a rise indistinguishable from zero is evidence that the readout does not condition on the prefix.*

The two tests are complementary and both run inside a single arm: Eq. (4) detects a readout *intervention* and is blind to an ignored prefix; Proposition 5 is the reverse. Together they are complete, in the following sense. Requiring Eq. (4) for *every* bounded $\mathcal{F}_t$-measurable h is the defining property of a conditional expectation, so Π satisfies the whole family at every t exactly when $(\tilde{P}_t)$ is a martingale in $(\mathcal{F}_t)$; and the one way to be a martingale without measuring anything is to be constant in expectation, which is the prefix-ignoring readout that Proposition 5 rejects. What a practitioner chooses is not whether to test the family but *which members*: the power sits entirely in h. We report two, $h \equiv 1$ and $h = P_t$ (Section 3), and Appendix T shows they catch different errors, with the second catching one that $h \equiv 1$ and Proposition 5 both miss.

3 Measuring the curve without moving it

The theory prescribes the protocol. **(1) Index at absolute token positions** $t \in \{t_1 < \cdots < t_m\}$, or at stopping times; never at a fraction of realised length or at a functional of the completed trace. **(2) Absorb**: a trace with $T \le t$ contributes δ_Y at t, so the population is the same at every index. **(3) Continue exactly**: sample K continuations of the prefix $c_{1:t}$ under the decoding law that produced the trace, with the same total cap M, so that $T = \min(T_{\text{EOS}}, M)$ is the same stopping time for base trace and continuations. With open weights we condition on the prefix's *token identifiers*, so continuation is exact by construction; with a closed model this needs a raw completion endpoint, and a chat endpoint offers at best an assistant-turn prefill whose validity is an empirical question (Section 4.3). **(4) Sample the base path**: a greedily decoded base trace with sampled continuations (Bigelow et al., 2024) is not a draw from the law the continuations follow, and Corollary 1 then does not apply; we test this variant separately. **(5) Report both curves** against their nulls, K flat at Acc and G rising from $\mathbb{E}\|P_0\|^2$.

Estimators. With K continuations per cell, the sample frequency $\hat{p}_t(a)$ is unbiased for $P_t(a)$ and $\hat{K}(t)$ follows by averaging with absorption. The agreement curve needs $\mathbb{E}\|P_t\|^2$, for which $\|\hat{p}_t\|^2$ is biased upward by $(1 - \|P_t\|^2)/K$; we use the U-statistic $\big(K\|\hat{p}_t\|^2 - 1\big)/(K-1)$, which is unbiased for any $K \ge 2$. Squared increments $\|\hat{p}_{t+1} - \hat{p}_t\|^2$ inherit twice the sampling variance and are corrected accordingly (Appendix D). Uncertainty is by cluster bootstrap over questions, and the paired readout effect is the per-trace difference between the forced and continued end-to-end changes. A completeness guard keeps only traces with every grid index absorbed or fully sampled (Appendix B).

The five internal checks. *Flat null* (Corollary 1): the slope of $\hat{K}$ on the grid, and its end-to-end change, must have confidence intervals containing zero. *Drift* (Eq. (4) with $h \equiv 1$): the number of grid intervals whose mean increment differs significantly from zero. *Conservation*: the increase of the bias-corrected $\mathbb{E}\|P\|^2$ across the grid must equal the sum of bias-corrected squared increments; this is Eq. (4) again at the weight $h = P_s$, and the gap between the two sides is $2\sum_s \mathbb{E}\langle P_s, \Delta_s\rangle$ exactly, per interval and per trace (Appendix A), so a failure can be localised rather than only reported. *Terminal agreement* (Proposition 5): the paired rise of $\hat{P}_t(Y)$ between a trace's shallowest and deepest sampled index must be positive. *Own-answer calibration* (Proposition 2): $\mathbb{E}[\hat{P}_t(Y) - \widehat{\|P_t\|^2}] = 0$ at every index, that is, a continuation reproduces the base trace's own answer as often as two continuations agree with each other. It is the conservation cross term taken to the terminal index rather than to the next grid point, so it spends every alive cell instead of one interval per trace. Three need a continuation arm, two do not, and they do not cover each other: Table 18 runs all five against 4 protocol errors realised as arms on the same traces, where 1 — conservation, which needs a continuation arm — catches all 4.

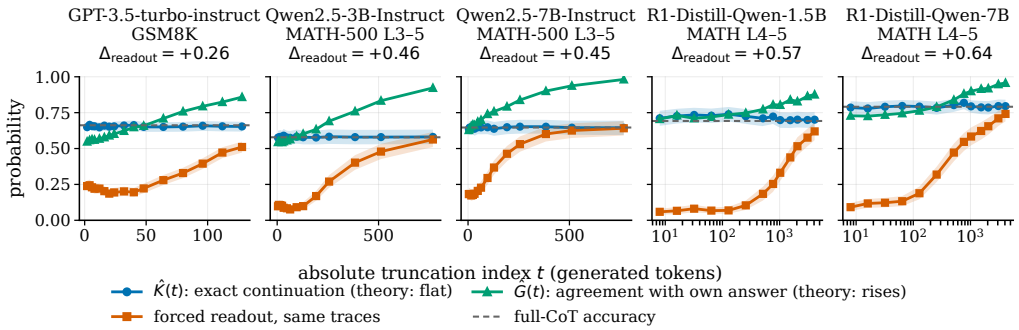

Figure 1: On every arm $\hat{K}(t)$ stays at full-CoT accuracy (dashed) while the forced-readout curve on the same traces and indices rises by Δ_{readout}, and $\hat{G}(t)$ rises as Proposition 2 requires. 95% cluster-bootstrap bands; log axis for the reasoning models.

Table 1: Main results. Slope of $\hat{K}$ on the standardised grid index [95% CI]; drift = intervals whose mean increment differs significantly from zero, continuation vs. forced, *before* the Benjamini–Hochberg correction across intervals that Section 4.4 applies (corrected counts in Table 5); Δ_{readout} = forced minus continuation end-to-end rise on the same traces [95% CI]; rise of $\hat{G}$.

arm	Acc	slope of $\hat{K}$ [95% CI]	drift	Δ_{readout} [95% CI]	$\hat{G}$ rise
GPT-3.5-turbo-instruct	0.662	−0.0004 [−0.0072, +0.0062]	0/15 vs 8/15	**+0.264** [+0.225, +0.301]	+0.310
Qwen2.5-3B-Instruct	0.578	+0.0005 [−0.0047, +0.0060]	1/13 vs 6/13	**+0.460** [+0.406, +0.512]	+0.378
Qwen2.5-7B-Instruct	0.646	+0.0010 [−0.0039, +0.0056]	5/13 vs 8/13	**+0.450** [+0.393, +0.505]	+0.353
R1-Distill-Qwen-1.5B	0.692	−0.0102 [−0.0190, −0.0018]	1/12 vs 8/12	**+0.570** [+0.505, +0.634]	+0.171
R1-Distill-Qwen-7B	0.791	+0.0012 [−0.0071, +0.0094]	1/12 vs 10/12	**+0.640** [+0.585, +0.693]	+0.229

4 Experiments

Arms. Table 3 lists the five arms. They span a closed model reached through a raw completion endpoint (text-level continuation), two instruction-tuned open models and two long-CoT reasoning models (token-exact continuation), and the datasets GSM8K (Cobbe et al., 2021) and the numeric-answer subset of MATH (Hendrycks et al., 2021), at difficulty levels chosen so that accuracy is interior. For each question we sample R base traces at temperature 1; for each base trace and each grid index $t < T$ we sample K exact continuations (the *continuation arm*) and K' forced readouts obtained by appending "*Given the above reasoning, the answer is*" and reading 16 tokens (the *forced arm*), so both curves are measured on the same traces at the same indices. Every open-weight arm also has a greedy-base-path sub-arm (temperature-0 base, temperature-1 continuations). Grids are geometric in absolute tokens and the generation cap M is one constant per arm. Answers are graded finish-aware: a continuation that hits the cap without an answer is $\perp$. In total 1 237 682 generations were recorded, 1 127 801 of them with vLLM (Kwon et al., 2023) (Appendix B). The five arms carry the main comparison; the checks below run on all 32 arms of Table 4, of which 31 were generated and one is derived.

4.1 The continuation curve is flat; the forced curve on the same traces is not

Fig. 1 and Table 1 carry the main result. Under exact continuation the correctness curve does not move: its slope is at most 0.0012 on four of the five arms with intervals containing zero, and on all five the end-to-end change is equivalent to zero at 0.03–0.05, an order of magnitude below the reported rises (Appendix P). The forced readout, on the *same* traces at the *same* indices, rises steeply on every arm, from 0.239 to 0.511 on GSM8K, and Table 1 gives the end-to-end readout effect for all five. The intervention term $K^{\text{force}}(t) - K(t)$ of Eq. (2) is negative throughout — −0.41 to −0.69 at the shallowest index, −0.00 to −0.14 at the deepest — so what a published figure shows is not a gap but the *closing* of one. That is what Δ_{readout} measures, the extra vertical extent the forced curve traverses over the same grid (Table 1), every interval excluding zero by a wide margin. A forced readout returning no number within its 16 tokens is $\perp$ and scored incorrect, as the literature scores it; neither that channel nor the readout's construction is what the effect measures (Appendix L: +0.27–+0.66 on traces parsing at both grid ends, and +0.633 to +0.537 in the worst case). On the

same samples $\hat{G}$ rises from an intercept of 0.548–0.729 toward one, as Proposition 2 requires: the two curves the literature conflates separate cleanly. Proposition 2 fixes $\hat{G}$'s endpoints but not its shape: with the intercept removed, the fraction of the remaining distance closed at half a trace's length runs 0.15–0.50 across 15 arms (Appendix O).

What is that vertical extent? Lanham et al. (2023)'s 0%-CoT point, the same readout with an empty prefix, answers it: across arms the intervention effect tracks the chain-of-thought gain $\mathrm{Acc} - \mathrm{Acc}_{0\%}$ with $r = 0.98$ (0.97–1.00 leaving any one arm out) and slope 0.83 (Appendix N), so truncating a chain and forcing an answer measures how much the chain was worth, not when the answer was decided. A correlation over arms we have already run is weak evidence for that reading, so we registered it as a prediction. Before generating anything we registered the rule on 6 arms with a 0.124 band, then ran Qwen2.5-14B-Instruct on MATH-500 under the 3B and 7B settings — twice the largest token-exact model above. Its chain is worth +0.597, so the registered prediction was +0.520 [+0.396, +0.643]; the measured effect is +0.633, inside the band and +0.113 from its centre. The flat null holds there too (slope -0.0005 [-0.0056, +0.0046], equivalent to no change at 0.03). The five arms pair each model with one dataset, so we ran a sixth, Qwen2.5-1.5B-Instruct, on 4 datasets including CommonsenseQA (Talmor et al., 2019) (Table 13). The flat null holds on every one while the intervention effect for that single model ranges from +0.03 to +0.50: the task, not only the model, sets the size. CommonsenseQA is the clean case — the chain is worth -0.024 there and the forced curve barely rises (+0.039 [+0.013, +0.064]).

4.2 What makes a published curve rise

A pipeline can commit the three errors together, so we also run the whole pipeline and attribute its rise: on the same traces, eight curves, one per combination of the three binary choices (Appendix J). Each is attested (forced readout in Lanham et al. (2023); Datta et al. (2026); Ballon et al. (2026), a fraction index in Lanham et al. (2023); Ballon et al. (2026); Nikulkov (2026), dropped finishers in any absolute-index pipeline without an absorption rule), but no one paper need make all three; the combination is the worst case, and it rises by +0.22 to +0.50 accuracy points. A sequential ledger would depend on the order of correction, so we report the exact Shapley attribution over the three choices (Fig. 2(a)): readout +0.24 to +0.56, survivorship -0.28 to -0.07, index +0.04 to +0.21, remainder the flat null. Survivorship, being negative, *masks* part of the readout rather than adding to it. The index term exceeds the covariance of Corollary 2 because under a forced readout a fraction index also moves the readout to the trace's own end; in the exact-continuation context the same swap is worth -0.02 to +0.01 (Appendix J). Each error is also measurable on its own and each matches the attribution: survivorship is the changing population exactly (under absorption $\pi_t \hat{S}(t) + (1 - \pi_t)\hat{F}(t) = \hat{K}(t)$), unpaired it costs -0.064 to -0.464 while the paired drop on traces alive at both ends is indistinguishable from zero on 13 of 15 arms; index leakage, estimated by $\hat{K}(\rho_f) - \hat{K}(t_1)$, is -0.012 to +0.015 at the pre-specified $f = 0.5$, none of the five tests survives Benjamini–Hochberg correction (Benjamini & Hochberg, 1995), staying within 2.3 points over all ten fractions; and the last-switch statistic of Datta et al. (2026) already reads 0.543–0.556 under a martingale whose answer is drawn only at the terminal step, so the published 35–70% bears on commitment only through its distance from that null (Appendices F, G and I).

4.3 Most chat endpoints silently ignore a trailing assistant prefix

Exact continuation needs token-level access or a raw completion endpoint; the cheap substitute on a chat API is to send the truncated reasoning back as a trailing assistant message. We tested whether 18 endpoints from four vendors continue such a message, with a prefix stopping mid-sentence and 24 samples each (Fig. 2(c), Appendix Q). Only 3 of 18 continue the prefix at least 95% of the time; 10 ignore it in every sample, 2 acknowledge and restart, and 3 are mixed. Vendor-documented modes (`prefix` at DeepSeek, `partial` at Alibaba) change nothing where they are accepted. The failure is silent: an ignored prefix gives the model answering from scratch at each "truncation point". Continuing a prefix is necessary rather than sufficient — re-encoding a decoded prefix, which is what an honoured prefill does, need not return the tokens the model emitted. On the same base traces and grid a Qwen2.5-7B arm continued from the *re-encoded* prefix tracks the token-exact arm to 0.014 at every index, passes the flat null and passes both checks that need no second arm — and fails conservation — gap -0.052 [-0.093, -0.008], and a paired shift of -0.0124 [-0.0210, -0.0037] in

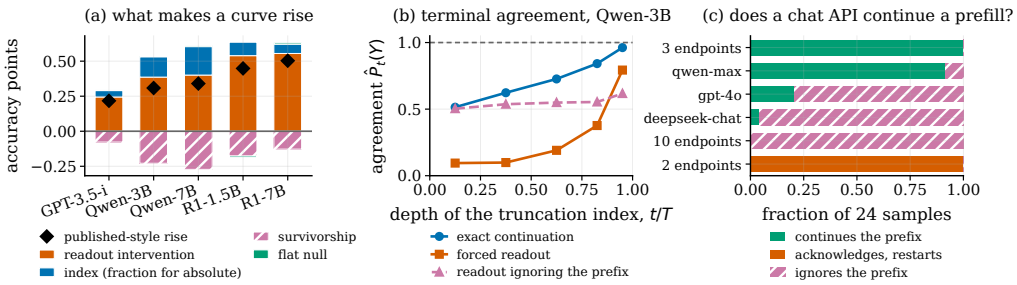

Figure 2: (a) A curve built the published way (forced readout, survivors only, fraction index) rises by +0.22 to +0.50 (diamonds); bars give the order-free Shapley attribution of that rise. (b) Terminal agreement (Proposition 5): near a trace's own end a real readout converges on that trace's answer, a prefix-ignoring one stays flat. (c) Whether a chat endpoint continues a trailing assistant-turn prefill, 24 samples each; endpoints with identical rates share a bar (rates per endpoint in Table 15).

the level statistic on the same cells (Appendix T). A curve can be conditioned on the wrong event and still pass every test a single-arm study can run; the check that catches it needs a continuation arm.

4.4 The internal checks, and where they fail

The drift test of Eq. (4) runs within each arm, and its per-interval rejections are a family of 10–15 tests, so an arm fails only when an interval survives Benjamini–Hochberg. All 5 forced arms then fail decisively (5–10 intervals survive), and all 5 main continuation arms pass. Terminal agreement (Proposition 5) passes on all 29 measurement arms of Table 4, paired depth rise +0.10–+0.66, every interval excluding zero. Conservation and own-answer calibration are one test per arm and are corrected across arms: of the 15 continuation arms, 1 survives Benjamini–Hochberg on conservation (the greedy-base 7B arm) and own-answer calibration rejects 2 (the greedy-base 3B and 7B arms). Neither is a surprise: a greedy base path is the arg-max trajectory, not a draw from P_t, so $\mathbb{E}[\hat{P}_t(Y)]$ exceeds $\mathbb{E}\|P_t\|^2$ there by construction. On every sampled-base arm the statistic is null to 0.021. No one check suffices, and Table 18 says how: run against 4 protocol errors realised as arms on the same traces, only conservation catches all 4, and it is the one that needs a continuation arm. To show what that buys we emulate a prefix-ignoring readout exactly: per cell a fresh sample of the same model, question and decoding law (Appendix K). It *passes* the flat null (+0.0034 [−0.0021, +0.0091]), a fresh draw at every index being a martingale in expectation; terminal agreement rejects it (+0.004 [−0.005, +0.014] against +0.293 for the real continuation, Fig. 2(b)).

One of the five main arms fails a check (every arm's verdicts are in Table 5): R1-Distill-Qwen-1.5B, whose flat-null slope excludes zero (−0.0102 [−0.0190, −0.0018]). The deviation is at most 0.047, its end-to-end change contains zero (−0.010 [−0.039, +0.020]), and it attenuates as the absorbed tail is removed, the interval covering zero once the grid stops at 1024 tokens (96% alive cells, Appendix R); the same-family 7B is flat at every depth.

5 Related work

Truncation curves and commitment. Turpin et al. (2023) showed CoT explanations can be unfaithful by perturbing the *prompt*, which the flat null leaves untouched since it does not truncate a trajectory; Lanham et al. (2023) introduced early answering and saw that the 0%-CoT point predicts the whole curve; Radhakrishnan et al. (2023) and Bentham et al. (2024) reuse those metrics. Ballon et al. (2026) plot accuracy at token percentiles; Datta et al. (2026), Scalena et al. (2026) and Zhang et al. (2026) locate a decision point from forced or probed readouts, and Zhang et al. (2025) read intermediate answers from hidden states, a probe Corollary 1 covers when it looks only at the past; Zhang et al. (2026) note that their index "is not an online stopping rule" without drawing the consequence. Chen et al. (2025) and Arcuschin et al. (2025) measure faithfulness on long-CoT models and without prompt perturbation, both on quantities the flat null does not govern; Garcia (2026) documents a different confound, answer position; Sivakumaran et al. (2026) train for the early-answering area. We add the null they must satisfy and decompose their departures.

Resampling and thought attribution. Bogdan et al. (2025) resample from sentence boundaries and argue qualitatively against forced answering as off-policy; Bigelow et al. (2024; 2026) sample exact continuations at absolute token positions, the protocol we advocate, but from a greedy base path and without aggregating against t; Ton et al. (2024) observe the flat curve for outcome reward models. None states the flat null, the survivor identity or the leakage covariance.

Martingale structure in LLM outputs. Falck et al. (2024) test whether in-context predictors are Bayesian through a martingale property; He et al. (2025) turn the orthogonality condition into a regression score; Kim et al. (2026) study drift on the answer simplex; Nikulkov (2026) proves the Doob property for outcome rewards and tests it binned by relative position, an instance of Corollary 2. They diagnose the model; we diagnose the measurement.

Early exit. Confidence- and stability-based early stopping (Yang et al., 2025; Liu & Wang, 2025; Fu et al., 2024; Huang et al., 2026; Ehab et al., 2026) and budget forcing (Muennighoff et al., 2025) rely on the same readouts, as do extended-thinking curves (Ghosal et al., 2025), which are K^{force} curves, and the no-thinking baseline of Ma et al. (2025), the 0%-CoT point of Appendix N; Corollary 1 says which stopping rules preserve the estimand (those looking only at the past), and Section 4.3 which readouts implement them.

6 Discussion

What a rising curve can still mean. The intervention curve $K^{\mathrm{force}}(t)$ is worth measuring: it is the accuracy of a model made to stop, which is what early-exit and budget-forcing systems deploy (Muennighoff et al., 2025). It cannot support the inference that the answer was "already decided" where it levels off. What can is the *shape* of the corrected agreement curve, which Proposition 2 leaves free: at half depth it has closed 0.15–0.50 of its distance to one, and that, measured on the corrected protocol (Appendix U), is what a commitment claim should be about. The distinction has teeth downstream: Sivakumaran et al. (2026) train to increase the area under an early-answering curve, and on our arms that area correlates with the intercept $\mathbb{E}\|P_0\|^2$ at 0.21–0.82 on the forced curve they optimise (Appendix P), so the objective rewards a concentrated answer prior.

Limitations. One arm fails the flat null by a margin we could not attribute. An earlier version of our own analysis graded each continuation without its prefix. That is right wherever a continuation restates an answer and wrong in the 0.1–2.8% of cells nearest a trace's end, and it was enough to manufacture violations of Proposition 2 on 6 arms. Own-answer calibration found it, and Appendix S reports it in full. After the readout correction of Appendix S the conservation gap of every sampled-base continuation arm is within 0.070 of zero, and only 1 of 15 survives correction across arms; but a residual inside the band an exact martingale gives at this design ([−0.068, +0.046]) is a bound, not a proof that the continuation law is exact; the same is true of the leakage bias. Terminal agreement has power only in proportion to $1 - \mathbb{E}\|P_0\|^2$, and own-answer calibration has none at all against a readout that ignores the prefix. Every closed-model endpoint was reached over one access path, so Section 4.3 reports endpoints as reached rather than vendor implementations (Appendix Q); the token-exact arms come from few model families on one serving stack (Appendix W). One limit is structural: across our tasks the chain's worth and the time of commitment are confounded, since a task where the chain is worth nothing is also one where an early-deciding model would produce a flat forced curve, so $\Delta_{\mathrm{readout}} \approx \mathrm{Acc} - \mathrm{Acc}_{0\%}$ is consistent with both readings. The flat null rules out K as evidence for either.

7 Conclusion

The truncated-CoT correctness curve is flat under exact continuation, by an elementary martingale argument. A published curve departs from it for three reasons we can name and measure, of which the readout dominates, at +0.24 to +0.56. Such a curve reports what the chain of thought is worth, not when the answer was fixed. Measuring one correctly means indexing at absolute positions or stopping times, absorbing finished traces, and continuing exactly, with $\mathbb{E}\|P_0\|^2$ and the 0%-CoT accuracy reported beside it. Separating a chain's worth from its time of commitment needs a task where the two come apart, which is the experiment we would run next.

AI USE STATEMENT

In this work, we used generative AI tools to assist in writing, editing, and debugging code, including code used to implement our methods and experiments. We did not use generative AI tools for any of the other tasks that require disclosure: in particular, we did not use them to develop theoretical models or conceptual frameworks, to formulate mathematical claims or assist in proofs, to propose or refine hypotheses, to design the research methodology or experiments, to assist with translation, to clean or reformat datasets, or to support data analysis or the interpretation of results. Nor did we use them to generate synthetic data. Two clarifications are needed here, because the subject matter of this paper is language-model output. The records we analyse are generations sampled from the language models under study, following the protocol of Section 3; they are the object of measurement, not a substitute for observations we could not make. The simulated belief martingale used to validate the estimators (Appendix E), and the null processes built on it, come from the numerical simulation in the released code rather than from a generative AI tool. Additionally, we used generative AI tools to edit the manuscript to improve readability and clarity, and to obtain simulated peer-review-style feedback on drafts, which informed revisions to the writing and presentation. All AI-assisted code was reviewed and tested for correctness by the authors, and all AI-suggested edits and feedback were reviewed and adopted at the authors' discretion; the scientific content, claims, and results are entirely the authors' own. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The supplementary archive holds the analysis code, the result file behind every table and figure, the question pools, the 18-endpoint audit replies, and all 1 237 682 generation records with their token counts, finish reasons and graded answers. Regenerating the paper's numbers from that archive alone reproduces every one of them; it needs no key, no GPU and no network. The records ship with their raw generated text replaced by the grader's verdict under both readings, since the text brings the archive past the upload limit, together with a seeded 2% sample of every pool that keeps its text so the grader can be checked against real generations. The complete pools with their raw text are held for release with the paper. Appendix B gives prompts, decoding parameters, grids and grading rules; Appendix D gives the estimators with their bias corrections and the bootstrap; Appendix A gives complete proofs; Appendix E describes the synthetic ground-truth validation of the estimators. Every measured quantity in the paper is generated by a single script from those result files, the only hand-written numerals being thresholds and analytic constants, so a re-analysis that changes a result changes the paper.

REFERENCES

Iván Arcuschin, Jett Janiak, Robert Krzyzanowski, Senthooran Rajamanoharan, Neel Nanda, and Arthur Conmy. Chain-of-Thought Reasoning In The Wild Is Not Always Faithful. *arXiv preprint arXiv:2503.08679*, 2025. ICML 2026.

Ned Augenblick and Matthew Rabin. Belief Movement, Uncertainty Reduction, and Rational Updating. *The Quarterly Journal of Economics*, 136(2):933–985, 2021.

Marthe Ballon, Brecht Verbeken, Vincent Ginis, and Andres Algaba. Probing the Trajectories of Reasoning Traces in Large Language Models. *arXiv preprint arXiv:2601.23163*, 2026.

Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289–300, 1995.

Oliver Bentham, Nathan Stringham, and Ana Marasović. Chain-of-Thought Unfaithfulness as Disguised Accuracy. *arXiv preprint arXiv:2402.14897*, 2024. TMLR 2024.

Eric Bigelow, Ari Holtzman, Hidenori Tanaka, and Tomer Ullman. Forking Paths in Neural Text Generation. *arXiv preprint arXiv:2412.07961*, 2024. ICLR 2025.

Eric Bigelow, Amir Zur, Satchel Grant, Tal Haklay, Can Rager, Owen Lewis, Thomas McGrath, Jack Merullo, Ekdeep Singh Lubana, and Atticus Geiger. Forking Fast: Efficiently Estimating Uncertainty Dynamics in Text Generation. *arXiv preprint arXiv:2608.19611*, 2026.

Paul C. Bogdan, Uzay Macar, Neel Nanda, and Arthur Conmy. Thought Anchors: Which LLM Reasoning Steps Matter? *arXiv preprint arXiv:2506.19143*, 2025.

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. Reasoning Models Don't Always Say What They Think. *arXiv preprint arXiv:2505.05410*, 2025.

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. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168*, 2021.

Ayan Datta, Zhixue Zhao, Bhuvanesh Verma, Radhika Mamidi, Mounika Marreddy, and Alexander Mehler. Large Language Models Decide Early and Explain Later. *arXiv preprint arXiv:2604.22266*, 2026.

Mohammed Ehab, Aymane El Gadarri, Vivek F. Farias, Adam Jozefiak, and Ciamac C. Moallemi. OS-Pruner: Pruning Chains-of-Thought of Reasoning Models via Optimal Stopping. *arXiv preprint arXiv:2607.11089*, 2026.

Fabian Falck, Ziyu Wang, and Chris Holmes. Is In-Context Learning in Large Language Models Bayesian? A Martingale Perspective. *arXiv preprint arXiv:2406.00793*, 2024. ICML 2024.

William Feller. *An Introduction to Probability Theory and Its Applications, Vol. I.* Wiley, 3rd edition, 1968.

Yichao Fu, Junda Chen, Siqi Zhu, Zheyu Fu, Zhongdongming Dai, Yonghao Zhuang, Yian Ma, Aurick Qiao, Tajana Rosing, Ion Stoica, and Hao Zhang. Efficiently Scaling LLM Reasoning with Certaindex. *arXiv preprint arXiv:2412.20993*, 2024.

Gabriel Garcia. The Last Word Often Wins: A Format Confound in Chain-of-Thought Corruption Studies. *arXiv preprint arXiv:2605.10799*, 2026.

Soumya Suvra Ghosal, Souradip Chakraborty, Avinash Reddy, Yifu Lu, Mengdi Wang, Dinesh Manocha, Furong Huang, Mohammad Ghavamzadeh, and Amrit Singh Bedi. Does Thinking More Always Help? Mirage of Test-Time Scaling in Reasoning Models. *arXiv preprint arXiv:2506.04210*, 2025. NeurIPS 2025.

Zhonghao He, Tianyi Qiu, Hirokazu Shirado, and Maarten Sap. Martingale Score: An Unsupervised Metric for Bayesian Rationality in LLM Reasoning. *arXiv preprint arXiv:2512.02914*, 2025. NeurIPS 2025.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring Mathematical Problem Solving With the MATH Dataset. In *NeurIPS Datasets and Benchmarks*, 2021.

Zixuan Huang, Xin Xia, Yuxi Ren, Jianbin Zheng, Xuanda Wang, Zhixia Zhang, Hongyan Xie, Songshi Liang, Zehao Chen, Xuefeng Xiao, Fuzhen Zhuang, Jianxin Li, Deqing Wang, and Yikun Ban. Does Your Reasoning Model Implicitly Know When to Stop Thinking? *arXiv preprint arXiv:2602.08354*, 2026.

SongEun Kim, Seungyoo Lee, Edwin Fong, Hyungi Lee, and Juho Lee. From Drift to Coherence: Stabilizing Beliefs in LLMs. *arXiv preprint arXiv:2606.17832*, 2026.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP)*, 2023.

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, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Measuring Faithfulness in Chain-of-Thought Reasoning. *arXiv preprint arXiv:2307.13702*, 2023. Anthropic tech report.

Paul Lévy. Sur certains processus stochastiques homogènes. *Compositio Mathematica*, 7:283–339, 1940.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's Verify Step by Step. *arXiv preprint arXiv:2305.20050*, 2023.

Xin Liu and Lu Wang. Answer Convergence as a Signal for Early Stopping in Reasoning. *arXiv preprint arXiv:2506.02536*, 2025.

Wenjie Ma, Jingxuan He, Charlie Snell, Tyler Griggs, Sewon Min, and Matei Zaharia. Reasoning Models Can Be Effective Without Thinking. *arXiv preprint arXiv:2504.09858*, 2025.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025.

Alex Nikulkov. Reward Models Are Secretly Value Functions: Temporally Coherent Reward Modeling. *arXiv preprint arXiv:2604.22981*, 2026.

Ansh Radhakrishnan, Karina Nguyen, Anna Chen, Carol Chen, Carson Denison, Danny Hernandez, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Sam McCandlish, Sheer El Showk, Tamera Lanham, Tim Maxwell, Venkatesa Chandrasekaran, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Question Decomposition Improves the Faithfulness of Model-Generated Reasoning. *arXiv preprint arXiv:2307.11768*, 2023. Anthropic report.

Daniel Scalena, Sara Candussio, Luca Bortolussi, Elisabetta Fersini, Malvina Nissim, and Gabriele Sarti. Beyond the Commitment Boundary: Probing Epiphenomenal Chain-of-Thought in Large Reasoning Models. *arXiv preprint arXiv:2606.13603*, 2026.

Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. Rewarding Progress: Scaling Automated Process Verifiers for LLM Reasoning. *arXiv preprint arXiv:2410.08146*, 2024. ICLR 2025.

Nithin Sivakumaran, Shoubin Yu, Hyunji Lee, Yue Zhang, Ali Payani, Mohit Bansal, and Elias Stengel-Eskin. Balancing Faithfulness and Performance in Reasoning via Multi-Listener Soft Execution. *arXiv preprint arXiv:2602.16154*, 2026.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. In *NAACL-HLT*, 2019. arXiv:1811.00937.

Jean-Francois Ton, Muhammad Faaiz Taufiq, and Yang Liu. Understanding Chain-of-Thought in LLMs through Information Theory. *arXiv preprint arXiv:2411.11984*, 2024. ICML 2025.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. *arXiv preprint arXiv:2305.04388*, 2023. NeurIPS 2023.

Peiyi Wang, Lei Li, Zhihong Shao, R. X. Xu, Damai Dai, Yifei Li, Deli Chen, Y. Wu, and Zhifang Sui. Math-Shepherd: Verify and Reinforce LLMs Step-by-step without Human Annotations. *arXiv preprint arXiv:2312.08935*, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *arXiv preprint arXiv:2201.11903*, 2022. NeurIPS 2022.

Chenxu Yang, Qingyi Si, Yongjie Duan, Zheliang Zhu, Chenyu Zhu, Qiaowei Li, Minghui Chen, Zheng Lin, and Weiping Wang. Dynamic Early Exit in Reasoning Models. *arXiv preprint arXiv:2504.15895*, 2025.

Anqi Zhang, Yulin Chen, Jane Pan, Chen Zhao, Aurojit Panda, Jinyang Li, and He He. Reasoning Models Know When They're Right: Probing Hidden States for Self-Verification. *arXiv preprint arXiv:2504.05419*, 2025.

Long Zhang, Wei-neng Chen, Feng-feng Wei, and Zi-bo Qin. When Does a Language Model Commit? A Finite-Answer Theory of Pre-Verbalization Commitment. *arXiv preprint arXiv:2605.06723*, 2026.

A PROOFS

Proposition 1. For each a, $P_t(a) = \mathbb{E}[\mathbf{1}\{Y = a\} \mid \mathcal{F}_t]$ with $\mathbf{1}\{Y = a\}$ bounded, so $(P_t(a))_t$ is a Doob martingale: $\mathbb{E}[P_{t+1}(a) \mid \mathcal{F}_t] = \mathbb{E}[\mathbb{E}[\mathbf{1}\{Y = a\} \mid \mathcal{F}_{t+1}] \mid \mathcal{F}_t] = P_t(a)$ by the tower property. Boundedness in the simplex is immediate. For $t \geq T$ the padded sequence $c_{1:t}$ contains the entire trajectory, so $Y = A(c)$ is $\mathcal{F}_t$-measurable and $P_t = \delta_Y$. Doob's convergence theorem is not needed since absorption occurs at the finite stopping time $T \leq M$.

Corollary 1. For deterministic t, $\mathbb{E}[P_t(a^\star)] = \mathbb{E}[\mathbb{E}[\mathbf{1}\{Y = a^\star\} \mid \mathcal{F}_t]] = \Pr[Y = a^\star]$. For a stopping time $\tau \leq M$, the optional stopping theorem for bounded martingales gives $\mathbb{E}[P_\tau(a^\star)] = \mathbb{E}[P_0(a^\star)]$. The examples listed (first sentence boundary, first occurrence of a token, first crossing of a probe computed from the prefix) are stopping times because each event $\{\tau \leq t\}$ is decided by $c_{1:t}$.

Proposition 2. Since Y is $\mathcal{F}_M$-measurable, $\mathbb{E}[P_t(Y) \mid \mathcal{F}_t] = \sum_a P_t(a) \Pr[Y = a \mid \mathcal{F}_t] = \sum_a P_t(a)^2$. The map $p \mapsto \|p\|^2$ is convex on the simplex, so conditional Jensen gives $\mathbb{E}[\|P_{t+1}\|^2 \mid \mathcal{F}_t] \geq \|\mathbb{E}[P_{t+1} \mid \mathcal{F}_t]\|^2 = \|P_t\|^2$. At $t = 0$, P_0 is the answer distribution of the model on the question (the prior over final answers), and $\|P_0\|^2$ is the probability that two independent trajectories agree. For $t \geq M$, $P_t = \delta_Y$ and $\|P_t\|^2 = 1$. The conservation identity follows from $\mathbb{E}[\|P_{s+1}\|^2 - \|P_s\|^2] = \mathbb{E}\|P_{s+1} - P_s\|^2 + 2\mathbb{E}\langle P_s, P_{s+1} - P_s\rangle$ and $\mathbb{E}\langle P_s, P_{s+1} - P_s\rangle = \mathbb{E}\langle P_s, \mathbb{E}[P_{s+1} - P_s \mid \mathcal{F}_s]\rangle = 0$.

Conservation is Eq. (4) with the weight $h = P_s$, and the estimator says so exactly. The last display makes the conservation identity a member of the orthogonality family rather than a separate fact: the gap between its two sides is

$$\underbrace{\mathbb{E}\|P_{t_m}\|^2 - \mathbb{E}\|P_{t_1}\|^2}_{\text{terminal gain}} \;-\; \underbrace{\sum_s \mathbb{E}\|P_{s+1} - P_s\|^2}_{\text{summed increments}} \;=\; 2\sum_s \mathbb{E}\big\langle P_s,\, \Delta_s \big\rangle, \tag{5}$$

and $\mathbb{E}\langle P_s, \Delta_s\rangle$ is Eq. (4) evaluated at the $\mathcal{F}_s$-measurable weight $h = P_s(\cdot)$, summed over the answer space. The drift test is the same condition at $h \equiv 1$. A weight that is itself the current belief puts the test's power where the process currently has mass, which is why the two can disagree on an arm.

With the estimators of Appendix D the decomposition is an algebraic identity per trace and per interval, not merely an identity in expectation. Writing $\widehat{\|P\|^2}$ for the U-statistic and $\widehat{\mathrm{var}}$ for the subtracted sampling variance, $\widehat{\mathrm{var}}(c) = \|\hat{p}\|^2 - \widehat{\|P\|^2}(c)$ exactly, so the bias-corrected increment is $\widehat{\|P_{s+1}\|^2} + \widehat{\|P_s\|^2} - 2\langle \hat{p}_s, \hat{p}_{s+1}\rangle$ and therefore

$$\widehat{\|P_{s+1}\|^2} - \widehat{\|P_s\|^2} - \widehat{\|\Delta_s\|^2} \;=\; 2\big(\langle \hat{p}_s, \hat{p}_{s+1}\rangle - \widehat{\|P_s\|^2}\big), \tag{6}$$

whose sum over s telescopes to the arm's measured gap with no residual. A failing arm can therefore be read interval by interval (Appendix S). Where the next index is absorbed, $\hat{p}_{s+1} = \delta_Y$ and the right-hand side becomes $2(\hat{P}_s(Y) - \widehat{\|P_s\|^2})$: the trace's own realised answer against the concentration of the continuations, which Proposition 2 forces to agree in expectation. That per-index form is the sharpest of the checks in Appendix T.

Proposition 3. $\{T > t\} \in \mathcal{F}_t$, so $\mathbb{E}[P_t(a^\star)\mathbf{1}\{T > t\}] = \mathbb{E}[\mathbf{1}\{Y = a^\star\}\mathbf{1}\{T > t\}]$ and $\mathbb{E}[P_t(a^\star)\mathbf{1}\{T \leq t\}] = \mathbb{E}[\mathbf{1}\{Y = a^\star\}\mathbf{1}\{T \leq t\}]$ by the tower property; summing and dividing by π_t and $1 - \pi_t$ gives $\mathrm{Acc} = \pi_t S(t) + (1 - \pi_t)F(t)$, with $F(t) = \mathbb{E}[\mathbf{1}\{Y = a^\star\} \mid T \leq t]$ because $P_t = \delta_Y$ on $\{T \leq t\}$.

Proposition 4 and Corollary 2. Given in the main text. For Corollary 2: $\rho_f = \lceil fT \rceil > s$ iff $fT > s$ iff $T > s/f$; since $s/f > s$ for $f < 1$, the event $\{T > s/f\}$ is not in $\mathcal{F}_s$ whenever $\Pr[s < T \leq s/f] > 0$. The sign statement is the observation that if, within a question, longer traces are less often correct, then at a fixed fraction the short (usually correct) traces are read at shallow absolute indices, before their belief has risen, and the long (usually incorrect) traces at deep indices, after their belief has fallen; both terms push the sum negative. At $f = 1$ every trace is read at its own T, where $P_T = \delta_Y$, and the mean is Acc; on a finite grid the read index is the nearest sampled point to fT, which for f near 1 is strictly below T whenever the grid is coarse there, so the vanishing is exact only for the ungridded index.

Corollary 3. This is a consequence of a modelling assumption, not of Proposition 1, and the corollary states it as such. *Assume* the running elicited answer A_s under the null is the sign of a centred symmetric walk with independent increments (in the binary case; for $|\mathcal{Y}| > 2$, of the difference between the leading two coordinates). A martingale on the simplex need not have such increments: they are neither symmetric nor identically distributed, and they concentrate as P_s does. Granting the assumption, the last sign change of a symmetric random walk of length n, divided by n, converges in law to the arcsine distribution with density $1/(\pi\sqrt{x(1-x)})$ (Lévy, 1940; Feller, 1968), whose mean is $\frac{1}{2}$ and which gives $\Pr[t^\star/T < 0.15] = \frac{2}{\pi}\arcsin\sqrt{0.15} \approx 0.253$. Because the assumption is doing the work, Appendix I replaces it with a simulated belief martingale read through each arm's own grid, and it is that null, not the arcsine value, that the paper tests against.

Proposition 5. Y is $\mathcal{F}_T$-measurable and $T \le M$ is finite, so Lévy's upward theorem gives $P_t(a) = \mathbb{E}[\mathbf{1}\{Y = a\} \mid \mathcal{F}_t] \to \mathbf{1}\{Y = a\}$ almost surely as $t \to T$; in particular $P_t(Y) \to 1$, and bounded convergence gives $\mathbb{E}[P_t(Y)] \to 1$ along the trace, from $\mathbb{E}[P_0(Y)] = \mathbb{E}\|P_0\|^2$. If instead the readout ignores the prefix and answers the question afresh under the same decoding law, then $\tilde{P}_t \overset{d}{=} P_0$ independently of t and of the trace, so $\mathbb{E}[\tilde{P}_t(Y)] = \mathbb{E}[P_0(Y)] = \mathbb{E}\|P_0\|^2$ for every t and the paired difference between any two depths has mean exactly zero. Only this direction is used: a paired rise indistinguishable from zero is evidence that the readout does not condition on the prefix. The converse is false in general—a readout could condition on the prefix in a way uncorrelated with the eventual answer—so a positive rise certifies dependence on the prefix, not correctness of the protocol, and it is the drift test of Eq. (4) that carries the other half.

B Experimental details

Prompts. Chat arms use the user message "*Solve this problem. Reason step by step but keep the reasoning tight and do not restate the problem. Finish with a final line of exactly the form* The answer is <number>. *Problem:* {problem}" through the model's chat template; the completion arm uses the same text followed by "*Solution:*". Forced readouts append "*Given the above reasoning, the answer is*" and generate 16 tokens; on the two reasoning models the suffix is appended inside the still-open reasoning block, as the truncation point falls there, and no closing tag is inserted. The answer extractor for continuations reads, in order, the number following "the answer is", the number inside `\boxed{}`, and—only for a generation that stopped on its own—the last number in the text. A generation that reaches the cap without one of the two explicit patterns is $\perp$, with the trailing-number fallback switched off for it, because on a truncated long trace that fallback returns an intermediate scratch value rather than an answer. The fallback resolves 19% of finished continuations, so disabling it everywhere would raise the $\perp$ rates of Table 2 rather than change any comparison, both curves being graded by the same rule. The extractor for forced readouts takes the number that immediately follows the suffix, terminated by punctuation, a closing delimiter or a word; a readout that opens with an expression or an operator ("6^3 in the denominator", "$58 - 40 =$") has not answered and is $\perp$. Table 2 gives the $\perp$ rate of every forced arm by index.

Table 2: Percentage of forced readouts with no parseable answer ($\perp$) at five grid indices (index: %), overall, and the same overall rate for the continuation arm on the same traces.

arm	first	quarter	middle	three-quarter	last	forced	continuation
GPT-3.5-turbo-instruct	2: 4	12: 6	32: 9	80: 14	128: 12	9	1
Qwen2.5-3B-Instruct	4: 25	24: 26	96: 44	256: 39	768: 31	35	1
Qwen2.5-7B-Instruct	4: 3	24: 4	96: 15	256: 24	768: 28	10	0
R1-Distill-Qwen-1.5B	8: 14	64: 25	512: 59	1536: 51	4096: 51	41	21
R1-Distill-Qwen-7B	8: 31	64: 23	512: 19	1536: 42	4096: 37	26	9

Datasets. GSM8K test, 350 questions with integer answers. MATH-500 restricted to numeric answers at levels 3–5 (223 questions) for the instruction-tuned models, and MATH at levels 4–5 (158 questions) for the reasoning models, chosen so that accuracy is interior. Ground truth is the dataset answer, compared numerically.

Decoding. Base traces and continuations use temperature 1.0 with no nucleus truncation. Continuations condition on the prefix's token identifiers (open weights, via the vLLM completion interface

with a list of ids) or on its text (completion API). The cap M is the total generation length for base traces; a continuation from index t may generate at most $M - t$ further tokens, so the stopping time $\min(T_{\text{EOS}}, M)$ is the same for every trace. Serving context is sized so that the longest prompt plus M always fits, since a server-side rejection of long prompts would silently exclude the longest questions.

Table 3: The five arms. Q questions, R base traces per question, K/K' continuations and forced readouts per cell, cap M in generated tokens. Traces after the completeness guard.

model	dataset	mechanism	grid (pts)	$Q \times R$	traces	K/K'	M
GPT-3.5-turbo-instruct	GSM8K	completion API, text	2–128 (16)	350×4	1 317	3/2	400
Qwen2.5-3B-Instruct	MATH-500 L3–5	vLLM, token ids	4–768 (14)	223×6	1 338	4/3	1536
Qwen2.5-7B-Instruct	MATH-500 L3–5	vLLM, token ids	4–768 (14)	223×6	1 338	4/3	1536
R1-Distill-Qwen-1.5B	MATH L4–5	vLLM, token ids	8–4096 (13)	158×3	474	3/3	8192
R1-Distill-Qwen-7B	MATH L4–5	vLLM, token ids	8–4096 (13)	158×3	474	3/3	8192

Grids. Indices count generated tokens of the model's own tokenizer, excluding the prompt.

- GPT-3.5-turbo-instruct: 2, 4, 6, 8, 12, 16, 20, 24, 32, 40, 48, 64, 80, 96, 112, 128.
- Qwen2.5-3B and 7B: 4, 8, 16, 24, 32, 48, 64, 96, 128, 192, 256, 384, 512, 768.
- R1-Distill 1.5B and 7B: 8, 16, 32, 64, 128, 256, 512, 768, 1024, 1536, 2048, 3072, 4096.

The *records* column of Table 3 counts generations written to a measurement pool, before the completeness guard below removes traces; the completion-API arm is the one whose two pools are not in the exact ratio K/K', a handful of requests having failed and been retried out.

Completeness guard. A trace enters the analysis only if every grid index is either absorbed ($t \geq T$) or has at least $K_{\min}$ samples ($K_{\min} = K$ for continuation arms, K' for forced arms, 2 for the completion arm's forced readouts). Without this guard the aggregate at different indices averages different trace populations, which is itself a survivorship artefact.

Host consistency. The Qwen2.5-7B continuation arm was generated on two physical GPUs of the same model class with the same software versions. We regenerated 1275 cells on the second host and compared answer distributions to the first at matched sample size: mean total-variation distance 0.2376 across hosts against 0.2412 within host, difference -0.0035, permutation $p = 0.8316$, minimum detectable difference 0.0275. The halves were merged. As a second check on pooling, the conservation statistic is recomputed within each host: +0.054 [-0.002, +0.110] on the 667 traces generated entirely on the first and +0.026 [-0.063, +0.112] on the 519 generated entirely on the second, against +0.039 [-0.005, +0.083] on the merged arm, three intervals containing zero and each other. (Before the readout correction of Appendix S the same three intervals were all negative and all excluded zero together, which is how we knew the gap was not the host change.)

C ARM INVENTORY

Table 3 lists the five main arms; the paper also reports checks that run on every arm we generated, and this is the full list. An *arm* is one (model, dataset, protocol) triple with its own grid, analysed as a unit. 15 exact-continuation arms (5 main, 4 greedy-base sub-arms, 4 one-model task arms, 1 held out, 1 second-family) and 14 forced-readout arms (5 main, 3 readout variants, 4 task arms, 1 held out, 1 second-family) are *measurement* arms; the 2 arms that exist to fail a check (Appendix T) and the derived prefix-ignoring arm of Appendix K bring the table to 32. Counts quoted in the main text refer to this table: terminal agreement is computed on all 29 measurement arms, the agreement-shape statistic on the 15 continuation arms, conservation on every continuation arm (after the readout correction of Appendix S, 1 of 15 survives correction across arms: the greedy-base 7B arm), own-answer calibration on every continuation arm (2 of 15 fail, the greedy-base 3B and 7B arms), the drift check on the 15 forced arms and on every continuation arm (Table 8 gives the greedy sub-arms), and survivorship on all 15 measurement continuation arms. Table 5 gives every arm's verdict on all five checks.

The 1 237 682 generations of Section 4 are the 1 107 670 records of the measurement pools below plus 130 012 in pools that are not an arm: the 13 700 base traces (sampled and greedy), the 72 028 readouts

of the 0%-CoT point (Appendix N), the 5 352-trace fresh pool behind the emulated arm (Appendix K), the 5 100 generations of the host-consistency probe (Appendix B: 1275 cells at $K = 4$), and the 33 832 generations of the serving-condition replicates (Appendix T).

Table 4: Every arm. "Records" counts generations in that arm's measurement pool (continuations or forced readouts); base pools are shared between the continuation and forced arms of a model and are counted once in 1 237 682. †cell assignments drawn from a separate pool of 5 352 fresh full traces (Appendix K), which is where those generations are counted, so this arm adds none of its own.

arm	model	dataset	protocol	grid	records
api_exact_cont	gpt-3.5-turbo-instruct	GSM8K	exact continuation	16	63 373
api_forced	gpt-3.5-turbo-instruct	GSM8K	forced readout, 16 tokens	16	42 308
qwen3b_cont	Qwen2.5-3B-Instruct	MATH L3–5	exact continuation	14	64 624
qwen3b_force	Qwen2.5-3B-Instruct	MATH L3–5	forced readout, 16 tokens	14	48 468
qwen3b_greedy_cont	Qwen2.5-3B-Instruct	MATH L3–5	continuation, greedy base	14	10 724
qwen7b_cont	Qwen2.5-7B-Instruct	MATH L3–5	exact continuation	14	55 092
qwen7b_force	Qwen2.5-7B-Instruct	MATH L3–5	forced readout, 16 tokens	14	41 319
qwen7b_greedy_cont	Qwen2.5-7B-Instruct	MATH L3–5	continuation, greedy base	14	9 300
r1d15b_cont	R1-Distill-Qwen-1.5B	MATH L4–5	exact continuation	13	15 762
r1d15b_force	R1-Distill-Qwen-1.5B	MATH L4–5	forced readout, 16 tokens	13	15 762
r1d7b_cont	R1-Distill-Qwen-7B	MATH L4–5	exact continuation	13	12 894
r1d7b_force	R1-Distill-Qwen-7B	MATH L4–5	forced readout, 16 tokens	13	12 894
r1d15b_greedy_cont	R1-Distill-Qwen-1.5B	MATH L4–5	continuation, greedy base	13	5 217
r1d7b_greedy_cont	R1-Distill-Qwen-7B	MATH L4–5	continuation, greedy base	13	4 236
r1d15b_forcec	R1-Distill-Qwen-1.5B	MATH L4–5	forced readout, block closed	13	15 762
r1d7b_forcec	R1-Distill-Qwen-7B	MATH L4–5	forced readout, block closed	13	12 894
r1d15b_forcelong	R1-Distill-Qwen-1.5B	MATH L4–5	forced readout, 48 tokens	13	10 508
qwen15b_math_cont	Qwen2.5-1.5B-Instruct	MATH L3–5	exact continuation	14	65 216
qwen15b_math_force	Qwen2.5-1.5B-Instruct	MATH L3–5	forced readout, 16 tokens	14	48 912
qwen15b_gsm_cont	Qwen2.5-1.5B-Instruct	GSM8K	exact continuation	12	55 832
qwen15b_gsm_force	Qwen2.5-1.5B-Instruct	GSM8K	forced readout, 16 tokens	12	41 874
qwen15b_csqa_cont	Qwen2.5-1.5B-Instruct	CommonsenseQA	exact continuation	11	33 588
qwen15b_csqa_force	Qwen2.5-1.5B-Instruct	CommonsenseQA	forced readout, 16 tokens	11	25 191
qwen3b_sham	Qwen2.5-3B-Instruct	MATH L3–5	prefix-ignoring (emulated)	14	64 624†
qwen14b_cont	Qwen2.5-14B-Instruct	MATH L3–5	exact continuation	14	58 400
qwen14b_force	Qwen2.5-14B-Instruct	MATH L3–5	forced readout, 16 tokens	14	43 800
qwen15b_arc_cont	Qwen2.5-1.5B-Instruct	ARC-Challenge	exact continuation	11	53 920
qwen15b_arc_force	Qwen2.5-1.5B-Instruct	ARC-Challenge	forced readout, 16 tokens	11	40 440
qwen7b_textcont	Qwen2.5-7B-Instruct	MATH L3–5	continuation, prefix re-encoded	14	55 092
qwen7b_textforce	Qwen2.5-7B-Instruct	MATH L3–5	forced readout, prefix re-encoded	14	41 319
llama8b_cont	Llama-3.1-8B-Instruct	MATH L3–5	exact continuation	14	58 828
llama8b_force	Llama-3.1-8B-Instruct	MATH L3–5	forced readout, 16 tokens	14	44 121

Table 5: The five internal checks of Section 3 on every arm. Flat null: the slope interval of $\hat{K}$ contains zero; conservation: the terminal gain and the summed corrected increments agree (interval of the difference contains zero); drift: no interval survives Benjamini–Hochberg (surviving count in parentheses); terminal agreement: the paired depth rise excludes zero above; own answer: $\hat{P}_t(Y) - \widehat{\|P_t\|^2}$ has an interval containing zero, which a forced readout is not required to satisfy and which is therefore not computed there ("–"). A forced readout is not a conditional expectation, so the first two have no reason to hold on a forced arm and the drift check is the one that detects it, with the qualification that where the chain is worth nothing there is almost no intervention to detect, which is why the ARC-Challenge forced arm passes the flat null; the emulated arm fails drift only on the intervals where absorbed cells enter (Appendix K). Among the 15 exact-continuation arms, 1 fails the flat null, 2 fail conservation, 2 fail drift, 0 fail terminal agreement and 2 fail own-answer calibration; the per-arm intervals here are uncorrected, and Appendix S gives the verdicts after correction across arms.

arm	model, dataset	protocol	flat null	conserv.	drift (BH)	term. agr.	own answer
api_exact_cont	GPT-3.5-turbo-instruct, GSM8K	exact continuation	pass	pass	pass (0)	pass	pass
api_forced	GPT-3.5-turbo-instruct, GSM8K	forced readout	**fail**	**fail**	**fail** (7)	pass	–
qwen3b_cont	Qwen2.5-3B, MATH-500	exact continuation	pass	pass	pass (0)	pass	pass
qwen3b_force	Qwen2.5-3B, MATH-500	forced readout	**fail**	**fail**	**fail** (5)	pass	–
qwen3b_greedy_cont	Qwen2.5-3B, MATH-500	greedy base	pass	pass	pass (0)	pass	**fail**
qwen7b_cont	Qwen2.5-7B, MATH-500	exact continuation	pass	pass	pass (0)	pass	pass
qwen7b_force	Qwen2.5-7B, MATH-500	forced readout	**fail**	**fail**	**fail** (7)	pass	–
qwen7b_greedy_cont	Qwen2.5-7B, MATH-500	greedy base	pass	**fail**	pass (0)	pass	**fail**
r1d15b_cont	R1-Distill-1.5B, MATH	exact continuation	**fail**	pass	pass (0)	pass	pass
r1d15b_force	R1-Distill-1.5B, MATH	forced readout	**fail**	**fail**	**fail** (8)	pass	–
r1d15b_forcec	R1-Distill-1.5B, MATH	forced, block closed	**fail**	**fail**	**fail** (9)	pass	–
r1d15b_forcelong	R1-Distill-1.5B, MATH	forced, 48 tokens	**fail**	**fail**	**fail** (8)	pass	–
r1d15b_greedy_cont	R1-Distill-1.5B, MATH	greedy base	pass	pass	**fail** (1)	pass	pass
r1d7b_cont	R1-Distill-7B, MATH	exact continuation	pass	pass	pass (0)	pass	pass
r1d7b_force	R1-Distill-7B, MATH	forced readout	**fail**	**fail**	**fail** (10)	pass	–
r1d7b_forcec	R1-Distill-7B, MATH	forced, block closed	**fail**	**fail**	**fail** (9)	pass	–
r1d7b_greedy_cont	R1-Distill-7B, MATH	greedy base	pass	pass	pass (0)	pass	pass
qwen15b_math_cont	Qwen2.5-1.5B, MATH-500	exact continuation	pass	pass	pass (0)	pass	pass
qwen15b_math_force	Qwen2.5-1.5B, MATH-500	forced readout	**fail**	**fail**	**fail** (5)	pass	–
qwen15b_gsm_cont	Qwen2.5-1.5B, GSM8K	exact continuation	pass	pass	pass (0)	pass	pass
qwen15b_gsm_force	Qwen2.5-1.5B, GSM8K	forced readout	**fail**	**fail**	**fail** (7)	pass	–
qwen15b_csqa_cont	Qwen2.5-1.5B, CommonsenseQA	exact continuation	pass	**fail**	pass (0)	pass	pass
qwen15b_csqa_force	Qwen2.5-1.5B, CommonsenseQA	forced readout	**fail**	**fail**	**fail** (3)	pass	–
qwen3b_sham	Qwen2.5-3B, MATH-500	prefix-ignoring (emulated)	pass	**fail**	**fail** (1)	**fail**	pass
qwen7b_textcont	Qwen2.5-7B, MATH-500	continuation, prefix re-encoded	pass	**fail**	pass (0)	pass	pass
qwen7b_textforce	Qwen2.5-7B, MATH-500	forced, prefix re-encoded	**fail**	**fail**	**fail** (8)	pass	–
qwen14b_cont	Qwen2.5-14B, MATH-500	exact continuation	pass	pass	pass (0)	pass	pass
qwen14b_force	Qwen2.5-14B, MATH-500	forced readout	**fail**	**fail**	**fail** (9)	pass	–
qwen15b_arc_cont	Qwen2.5-1.5B, ARC-Challenge	exact continuation	pass	pass	pass (0)	pass	pass
qwen15b_arc_force	Qwen2.5-1.5B, ARC-Challenge	forced readout	pass	**fail**	**fail** (1)	pass	–
llama8b_cont	Llama-3.1-8B, MATH-500	exact continuation	pass	pass	**fail** (1)	pass	pass
llama8b_force	Llama-3.1-8B, MATH-500	forced readout	**fail**	**fail**	**fail** (6)	pass	–

D ESTIMATORS

With K continuations of a cell and empirical frequencies $\hat{p}$, $\mathbb{E}\|\hat{p}\|^2 = \|P\|^2 + (1 - \|P\|^2)/K$, so $\widehat{\|P\|^2} = (K\|\hat{p}\|^2 - 1)/(K - 1)$ is unbiased. For an increment between two independent cells with K_1, K_2 samples, $\mathbb{E}\|\hat{p}_2 - \hat{p}_1\|^2 = \|P_2 - P_1\|^2 + (1 - \|P_1\|^2)/K_1 + (1 - \|P_2\|^2)/K_2$, and we subtract the unbiased estimates of the two variance terms. On synthetic data (Appendix E) the uncorrected sum of squared increments exceeds the terminal gain by a factor of 2.5 at $K = 8$; the corrected sum matches it. The correction is exact in expectation for any $K \geq 2$, and it holds at the smallest cell sizes we use: on the same synthetic martingale over 20 seeds, the mean of the difference is +0.008 (s.e. 0.006) at $K = 3$, -0.005 (0.006) at $K = 4$ and -0.006 (0.003) at $K = 8$, with 1, 0 and 0 of 20 seeds failing the check at the 5% level. Finite K widens the interval of the check and moves its centre by at most a few thousandths. Absorbed cells have $\hat{p} = \delta_Y$ exactly and contribute no sampling variance.

Confidence intervals are percentile intervals from 4 000 cluster-bootstrap resamples of questions, applied to the per-trace quantity in question. The flat-null slope is the least-squares slope of the mean correctness curve on the grid index standardised to zero mean and unit variance (so that the statistic is comparable across arms with different grids), bootstrapped over questions. The paired readout

effect is $\left[\hat{p}^{\text{force}}_{t_m}(a^\star) - \hat{p}^{\text{force}}_{t_1}(a^\star)\right] - \left[\hat{p}^{\text{cont}}_{t_m}(a^\star) - \hat{p}^{\text{cont}}_{t_1}(a^\star)\right]$ averaged over traces present in both arms. The drift count is the number of adjacent grid intervals whose mean per-trace increment has a bootstrap interval excluding zero. The index-leakage permutation test shuffles T among the traces of each question 1 000 times and recomputes the paired statistic with each trace read at $\lceil fT' \rceil$ for its permuted length T'; the reported p is two-sided about the null mean.

E SYNTHETIC VALIDATION

Before any language-model data were collected, the estimators were run on a simulated belief martingale with known ground truth: a Dirichlet answer prior per question, a Bayesian posterior updated by noisy observations of the eventual answer until it exceeds a threshold, absorption at a terminal draw, and $K = 8$ continuations per cell. 3 seeds of 400 questions $\times$ 6 traces were run. The estimated correctness curve deviates from accuracy by at most 0.013, within Monte-Carlo error and with slope intervals containing zero. The bias-corrected conservation identity holds (terminal gain 0.58 against summed increments 0.55–0.57, intervals containing zero) while the uncorrected sum is 1.42–1.46. The fraction-indexed curve of Corollary 2 shows a significant negative paired leakage of -0.016 to -0.028 at interior fractions and a manufactured rise of +0.013 to +0.027 over the second half of the trace, with no learning anywhere in the process, and both shrink toward zero as the length–correctness coupling is swept toward zero. The last-switch fraction of this process, whose answer is drawn only at the terminal step, has mean post-decision fraction 0.57–0.58 over these seeds, with 0.30–0.31 of its mass below 0.15 against the arcsine prediction 0.25 at this trace length. Appendix I recomputes the same quantity on the separate $1\,800$-trace simulation used for the null there and gets 0.589; the difference is which simulation, not a different process. The per-cell estimators are unbiased to within $|z| \leq 1.40$ on 16 013 cells.

F SURVIVORSHIP ON ALL ARMS

Table 6: Survivor curve $S(t)$ drop across the widest grid range with at least 30 survivors (unpaired), the drop on the same traces alive at both ends (paired, with 95% CI), and the maximum reconstruction error $|\pi_t S(t) + (1 - \pi_t)F(t) - \text{Acc}|$ over the grid.

arm	unpaired drop	paired drop [95% CI]	max reconstruction error
GPT-3.5-turbo-instruct, GSM8K	−0.064	−0.002 [−0.029, +0.025]	0.014
Llama-3.1-8B, MATH-500	−0.319	−0.053 [−0.083, −0.026]	0.018
Qwen2.5-1.5B, ARC-Challenge	−0.106	+0.017 [−0.029, +0.062]	0.018
Qwen2.5-1.5B, CommonsenseQA	−0.191	−0.048 [−0.134, +0.035]	0.013
Qwen2.5-1.5B, GSM8K	−0.316	−0.027 [−0.089, +0.035]	0.012
Qwen2.5-1.5B, MATH-500	−0.206	−0.018 [−0.061, +0.025]	0.006
Qwen2.5-14B, MATH-500	−0.388	−0.013 [−0.106, +0.081]	0.010
Qwen2.5-3B, MATH-500	−0.362	−0.022 [−0.068, +0.022]	0.011
Qwen2.5-3B, greedy base	−0.288	+0.083 [+0.006, +0.167]	0.050
Qwen2.5-7B, MATH-500	−0.464	+0.000 [−0.092, +0.083]	0.010
Qwen2.5-7B, greedy base	−0.376	+0.012 [−0.098, +0.134]	0.031
R1-Distill-1.5B, MATH	−0.241	−0.029 [−0.083, +0.019]	0.047
R1-Distill-1.5B, greedy base	−0.219	+0.010 [−0.086, +0.106]	0.078
R1-Distill-7B, MATH	−0.125	−0.019 [−0.086, +0.045]	0.027
R1-Distill-7B, greedy base	−0.213	−0.122 [−0.267, +0.011]	0.042

G INDEX LEAKAGE AT EVERY FRACTION

Table 7: Paired index-leakage statistic $\hat{K}(\rho_f) - \hat{K}(t_1)$ at each fraction f of realised trace length, with the minimum detectable effect (MDE, 95%) at $f = 0.5$. *: 95% cluster-bootstrap interval excludes zero. Only $f = 0.5$ on the five sampled-base arms was pre-specified; the greedy sub-arms are outside Corollary 1 and are shown for completeness.

arm	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	MDE
GPT-3.5-turbo-instruct	−0.002	+0.002	−0.001	+0.006	−0.004	−0.002	+0.003	+0.003	+0.005	+0.005	0.021
Qwen2.5-3B-Instruct	−0.008	−0.006	−0.000	+0.002	+0.006	+0.005	+0.001	+0.008	+0.006	+0.007	0.015
Qwen2.5-7B-Instruct	+0.014*	+0.008	+0.004	+0.010	+0.015*	+0.011	+0.012	+0.010	+0.006	+0.006	0.015
R1-Distill-Qwen-1.5B	+0.018	−0.005	−0.006	−0.004	−0.012	−0.015	−0.013	−0.007	−0.002	−0.004	0.029
R1-Distill-Qwen-7B	+0.005	+0.009	−0.023	+0.001	+0.011	+0.010	+0.012	+0.009	+0.011	+0.011	0.028
Qwen2.5-3B, greedy base	−0.006	+0.000	+0.009	+0.001	+0.007	+0.010	+0.020	+0.021	+0.029	+0.037	0.037
Qwen2.5-7B, greedy base	+0.009	−0.007	−0.010	−0.017	−0.008	−0.003	+0.007	+0.007	+0.010	+0.009	0.034
R1-Distill-1.5B, greedy base	+0.034	+0.038	+0.030	+0.034	+0.049*	+0.053*	+0.055*	+0.061*	+0.063*	+0.061*	0.051
R1-Distill-7B, greedy base	−0.030	−0.034	−0.032	−0.015	−0.042	−0.030	−0.034	−0.038	−0.040	−0.040	0.048

At $f \geq 0.9$ the grid point nearest fT can lie at or beyond the trace's own end, where the absorbed cell returns realised correctness; the negative values there on Qwen2.5-7B and R1-Distill-7B (greedy) are the leakage of Corollary 2 at an index that reveals almost the whole future, and their sign follows the within-question length–correctness correlation as the corollary says.

H GREEDY BASE-PATH SUB-ARMS

Table 8: Continuation arms whose base trace was decoded greedily (temperature 0) with sampled continuations, the protocol of Bigelow et al. (2024). One trace per question. "After BH" is the number of grid intervals surviving Benjamini–Hochberg within the arm, the decision rule of Section 4.4; R1-Distill-1.5B's single surviving interval makes it the one sub-arm that fails the drift check.

arm	traces	Acc	slope of $\hat{K}$ [95% CI]	drift (raw)	after BH	verdict
Qwen2.5-3B-Instruct	223	0.610	+0.0107 [−0.0039, +0.0244]	0/13	0	**pass**
Qwen2.5-7B-Instruct	223	0.668	+0.0063 [−0.0059, +0.0187]	0/13	0	**pass**
R1-Distill-Qwen-1.5B	158	0.677	+0.0050 [−0.0091, +0.0197]	1/12	1	**fail**
R1-Distill-Qwen-7B	158	0.759	−0.0004 [−0.0167, +0.0150]	2/12	0	**pass**

I LAST-SWITCH NULL

Table 9: Mean post-decision fraction $1 - t^\star/T$ of the forced-readout answer, against the belief-martingale null of Appendix E read through the same grid, and the same null observed at every step (the continuous arcsine law gives $\frac{1}{2}$).

	GPT-3.5-i	Qwen-3B	Qwen-7B	R1-1.5B	R1-7B
observed	0.227	0.110	0.209	0.362	0.373
null, same grid	0.548	0.556	0.552	0.551	0.543
null, continuous	0.589	0.589	0.589	0.589	0.589
KS(observed, null)	0.463	0.563	0.428	0.313	0.302

The belief martingale of Appendix E, a Bayesian posterior that concentrates on an answer drawn only at the terminal step, is simulated for 1 800 traces. It is a null for *commitment before the terminal draw*, not the symmetric walk of Corollary 3, and its increments are not symmetric; we use it because it is the process the estimators were validated on. For each arm, the real grid $\{t_i\}$ is mapped to the simulation by $t_i \mapsto \mathrm{round}(t_i \cdot \tilde{m}/m)$, where m and $\tilde{m}$ are the median trace lengths of the arm and of the simulation, so that the simulated process is observed at the same fractions of trace length as the real one; the elicited answer at each grid point is the argmax of the simulated belief (or the realised answer if absorbed), and $t^\star$ is computed exactly as for the real arm, including the terminal point (T, Y). "Null, continuous" in Table 9 is the same process observed at every step. The gap between the continuous null and the arcsine value $\frac{1}{2}$ is mostly a finite-length effect rather than a property of the process: a symmetric ± 1 walk with the same trace lengths (median 55 steps) has mean post-decision

fraction 0.574 and 0.34 of its mass below 0.15, against 0.589 and 0.32 for the belief martingale, both computed the same way on the same $1\,800$ traces.

Our arms sit below the null, not at it. The point of Corollary 3 is that a published post-decision fraction has to be read against a null, not against zero; it is not that our arms reproduce the null. They do not: observed fractions of 0.11–0.37 against a grid-matched null of 0.543–0.556 put our own models *earlier* than a process with no commitment at all, by Kolmogorov–Smirnov distances of 0.30–0.56. Read literally that says these models settle sooner than a martingale that draws its answer only at the end — which is a statement about an argmax-of-belief statistic computed on a coarse grid, not about commitment, and it is exactly why the paper's positive claim rests on the shape of $\hat{G}$ (Appendix O) rather than on $t^\star$.

J THE PUBLISHED-PIPELINE DECOMPOSITION

Four curves on the same traces, differing in one choice at a time: R_1 forced readout, survivors only, index $\rho_f = \lceil fT \rceil$ discretised onto the grid; R_2 the same with an absolute index; R_3 the same with absorption; R_4 exact continuation with absorption. Each rise is end-to-end on its own axis. Writing the three binary choices as a coalition game whose value is the rise of the corresponding protocol, the Shapley value of each choice averages its marginal contribution over all six orders; all eight protocols are measured, so the attribution is exact rather than sampled, and the four terms sum to R_1 by construction. Uncertainty is a cluster bootstrap over questions with the same resampled questions feeding all eight curves, so the components are paired. The index term is not the covariance of Corollary 2 alone: under a forced readout, a fraction index reads the trace at its own end when $f = 1$, where the forced curve has closed most of its gap to the continuation curve, so the swap from an absolute grid to a fraction grid changes what the forced readout is applied to and not only which future the index reveals. The swap in the exact-continuation context, $V(0,0,1) - V(0,0,0)$, isolates the covariance term and is -0.02 to +0.01 across the five arms (at most 1.6 points), consistent with the paired statistic of Appendix G.

A published curve is usually read for a *commitment point*, the index at which it has reached most of its rise. On these arms the published combination places that point (at 90% of the rise) between 70 and 100% of the trace. Under exact continuation the total rise is inside the equivalence margin of Appendix P, so there is no rise for such a point to sit on and any value read off is sampling noise—which is the practical content of Corollary 1.

Table 10: Published-style rise and its Shapley attribution, with 95% cluster-bootstrap intervals. The four attribution columns sum to the published rise exactly, row by row; the last of them is the exact-continuation curve's own end-to-end change, which the flat null puts at zero and which is therefore small and of either sign. It is printed so that the row can be added up. The index column is the fraction-for-absolute swap averaged over the four readout–population contexts; in the exact-continuation context alone the swap is +0.002, +0.010, -0.016, -0.012 and -0.004, the statistic of Corollary 2 that Appendix G measures at each fraction.

arm	published rise	readout	survivorship	index	flat null
GPT-3.5-turbo-instruct	+0.217 [+0.177, +0.258]	+0.244 [+0.208, +0.279]	−0.077 [−0.094, −0.060]	+0.046 [+0.030, +0.060]	+0.004
Qwen2.5-3B-Instruct	+0.308 [+0.262, +0.359]	+0.386 [+0.338, +0.432]	−0.228 [−0.262, −0.194]	+0.145 [+0.116, +0.176]	+0.006
Qwen2.5-7B-Instruct	+0.340 [+0.283, +0.401]	+0.401 [+0.346, +0.458]	−0.272 [−0.321, −0.220]	+0.203 [+0.158, +0.244]	+0.008
R1-Distill-Qwen-1.5B	+0.448 [+0.385, +0.512]	+0.539 [+0.484, +0.599]	−0.177 [−0.212, −0.143]	+0.096 [+0.068, +0.123]	−0.010
R1-Distill-Qwen-7B	+0.503 [+0.439, +0.565]	+0.555 [+0.501, +0.613]	−0.128 [−0.171, −0.085]	+0.066 [+0.030, +0.105]	+0.010

K TERMINAL AGREEMENT, AND THE EMULATED IGNORED PREFIX

For every unabsorbed cell we record $\hat{P}_t(Y)$, the sampled frequency of the base trace's own realised answer, against the relative depth t/T. The paired statistic takes, within each trace, the deepest sampled index with $t/T \geq 0.9$ and the shallowest with $t/T \leq 0.25$, and averages the difference; pairing removes the composition difference between depth bins, which otherwise contributes a small positive rise because deep bins over-represent short traces.

The emulated ignored-prefix readout answers each cell with a fresh sample of the same model on the same question under the same decoding law and cap, drawn without replacement from a pool of 24

fresh traces per question generated for this purpose. This is exactly what an endpoint that drops a trailing assistant message returns at every "truncation point". Its curve passes the flat null and fails both terminal agreement (Section 4.4) and conservation (0.314 against -0.144, difference CI [+0.393, +0.525]): a fresh draw at every index carries quadratic variation that no terminal gain can account for, so the two checks that need only one arm catch it from opposite directions. The drift test flags only the two intervals in which absorbed cells enter the average, whose values come from the analyst's absorption convention and not from the readout; below that region it finds nothing, as Eq. (4) predicts for pseudo-beliefs that are constant in expectation.

Table 11: Terminal agreement: mean $\hat{P}_t(Y)$ in the deepest depth bin, and the paired within-trace rise from the shallowest to the deepest sampled index, with 95% intervals.

arm	exact continuation		forced readout	
	deepest bin	paired rise	deepest bin	paired rise
GPT-3.5-turbo-instruct	0.992	+0.300 [+0.274, +0.327]	0.981	+0.441 [+0.404, +0.476]
Qwen2.5-3B-Instruct	0.962	+0.293 [+0.260, +0.328]	0.793	+0.356 [+0.319, +0.394]
Qwen2.5-7B-Instruct	0.974	+0.260 [+0.226, +0.293]	0.844	+0.391 [+0.343, +0.439]
R1-Distill-Qwen-1.5B	0.981	+0.141 [+0.108, +0.174]	0.871	+0.627 [+0.578, +0.672]
R1-Distill-Qwen-7B	0.955	+0.181 [+0.144, +0.222]	0.857	+0.653 [+0.610, +0.698]
emulated ignored prefix (Qwen2.5-3B): deepest bin 0.617, paired rise +0.004 [−0.005, +0.014]				

L THE READOUT EFFECT WITHOUT THE PARSE FAILURES

A forced readout that produces no answer inside its window is $\perp$ and is scored incorrect, exactly as the literature scores it. It happens often (9–41% of forced readouts against at most 21% of continuations, Table 2), so part of the measured intervention effect could in principle be the model failing to answer rather than answering wrongly. The restrictions below leave the effect at +0.27–+0.66 on traces parsing at both grid ends and +0.26–+0.62 with $\perp$ dropped cell-wise. Two restrictions remove that channel. *Parseable at both ends*: keep only traces whose forced readout parses at the first and the last grid index, and recompute the paired contrast on them. *Bottom dropped*: renormalise every forced cell over its parseable draws only, so $\perp$ never enters an estimate. On the five main arms both leave the effect at least as large as the headline; across all 15 arms with a forced counterpart the restriction to traces parsing at both ends keeps the sign everywhere and at least 85% of the effect, the largest reduction being Qwen2.5-14B (+0.633 to +0.537).

Table 12: Paired readout effect with 95% cluster-bootstrap intervals: as reported, on traces whose forced readout parses at both grid ends, and with $\perp$ draws dropped cell-wise.

arm	as reported	parseable at both ends	$\perp$ dropped
GPT-3.5-turbo-instruct	+0.264 [+0.225, +0.301]	+0.273 [+0.234, +0.311] ($n = 1\,266$)	+0.264 [+0.225, +0.301]
Qwen2.5-3B-Instruct	+0.460 [+0.406, +0.512]	+0.469 [+0.415, +0.522] ($n = 1\,207$)	+0.450 [+0.394, +0.502]
Qwen2.5-7B-Instruct	+0.450 [+0.393, +0.505]	+0.450 [+0.393, +0.504] ($n = 1\,326$)	+0.450 [+0.393, +0.504]
R1-Distill-Qwen-1.5B	+0.570 [+0.505, +0.634]	+0.663 [+0.606, +0.719] ($n = 402$)	+0.594 [+0.532, +0.658]
R1-Distill-Qwen-7B	+0.640 [+0.585, +0.693]	+0.662 [+0.606, +0.715] ($n = 423$)	+0.620 [+0.562, +0.674]

Readout variants on a reasoning model. The truncation index of a long-CoT model falls inside an open reasoning block, so the literature's suffix asks for an answer mid-thought. Two variants test whether that is what the effect measures. *Closed block*: insert "\n</think>\n\n" before the suffix, which is also what deployed budget forcing does (Muennighoff et al., 2025). *Longer window*: leave the block open and read 48 tokens instead of 16. Closing the block cuts the $\perp$ rate from 41% to 7% on R1-Distill-1.5B and from 26% to 13% on R1-Distill-7B, and the paired effect moves to +0.633 [+0.573, +0.691] and +0.617 [+0.558, +0.673] respectively, against +0.570 [+0.505, +0.634] and +0.640 [+0.585, +0.693] for the naive readout—up on one model, down on the other, both intervals covering the naive value. The longer window changes nothing at all ($\perp$ 41%, effect +0.573 [+0.509, +0.637]). The parse failures were not creating the measured intervention; on the 1.5B they were suppressing it.

M ONE MODEL, 4 DATASETS

Table 13: Qwen2.5-1.5B-Instruct on 4 datasets, generated and analysed exactly like the main arms. $\hat{G}(0)$ is the concentration of the answer prior; CommonsenseQA and ARC-Challenge have letter answer spaces, MATH-500 and GSM8K open numeric ones.

dataset	traces	Acc	$\hat{G}(0)$	slope of $\hat{K}$ [95% CI]	Δ_{readout} [95% CI]	$\hat{G}$ rise
MATH-500 L3–5	1 338	0.339	0.280	−0.0019 [−0.0083, +0.0043]	**+0.243** [+0.198, +0.286]	+0.610
GSM8K	1 400	0.609	0.517	−0.0019 [−0.0085, +0.0048]	**+0.504** [+0.465, +0.542]	+0.458
CommonsenseQA	1 000	0.633	0.661	+0.0014 [−0.0062, +0.0092]	**+0.039** [+0.013, +0.064]	+0.328
ARC-Challenge	1 500	0.661	0.672	+0.0035 [−0.0019, +0.0089]	**+0.028** [−0.001, +0.059]	+0.288

Grids, caps and sample sizes follow the corresponding main arms; the two multiple-choice prompts ask for a final line "*The answer is* ⟨*letter*⟩" and the extractor reads a letter in `A`–`E` by the same priority order as the numeric one. The 4 arms share a model and differ only in the task, so the spread of Δ_{readout} across them, wider than the spread across the five main models, shows that the size of the readout intervention is a property of the question and its answer space, not a measure of how early the model decided. The flat null is what does not move.

The two non-mathematical tasks say the same thing twice, independently: on CommonsenseQA the chain is worth −0.024 and the curve rises +0.039, on ARC-Challenge it is worth −0.015 and the curve rises +0.028. ARC was added to fill the middle of the regression of Appendix N and did not, landing beside CommonsenseQA instead, so what it buys is a replication of the control rather than a new point: where the chain carries no information, the protocol that is supposed to detect early commitment detects almost nothing, on two unrelated tasks.

Conservation on these arms. All 4 pass the conservation check of Section 4.4 after the correction of Appendix S on the corrected readout of Appendix S: MATH-500 [−0.018, +0.084], GSM8K [−0.035, +0.060], ARC-Challenge [−0.059, +0.038], and CommonsenseQA [−0.138, −0.003], the only one whose own interval excludes zero and which does not survive correction across arms. Before that correction three of the four failed, by −0.408 on CommonsenseQA and −0.340 on ARC-Challenge. Those are the two tasks with the smallest answer spaces, where a deficit of a fixed rate is largest in absolute terms, which is why the artefact was visible here first. Own-answer calibration passes on all 4. The flat null, which is what these arms are for, holds on all 4.

N THE 0%-COT READOUT

The same forced readout with an empty prefix is Lanham et al. (2023)'s 0%-CoT point: the model is asked for an answer with no reasoning at all, under the same prompt, decoding law and 16-token window as every other forced cell. It bounds what a truncation curve can traverse, because a forced readout at index t differs from the 0% point only in how much of the chain it has been given.

Table 14: Accuracy with no chain of thought, with the full chain, the difference between them, and the measured readout effect. Intervals are cluster bootstraps over questions.

arm	$\text{Acc}_{0\%}$ [95% CI]	Acc	CoT gain	Δ_{readout}
GPT-3.5-turbo-instruct, GSM8K	0.246 [+0.207, +0.285]	0.662	+0.416	+0.264
Qwen2.5-3B, MATH-500 L3–5	0.057 [+0.038, +0.078]	0.578	+0.521	+0.460
Llama-3.1-8B, MATH-500 L3–5	0.076 [+0.053, +0.101]	0.359	+0.284	+0.289
Qwen2.5-7B, MATH-500 L3–5	0.183 [+0.143, +0.225]	0.646	+0.463	+0.450
R1-Distill-1.5B, MATH L4–5	0.067 [+0.047, +0.089]	0.692	+0.625	+0.570
R1-Distill-7B, MATH L4–5	0.082 [+0.060, +0.108]	0.791	+0.709	+0.640
Qwen2.5-1.5B, MATH-500 L3–5	0.088 [+0.065, +0.114]	0.339	+0.250	+0.243
Qwen2.5-1.5B, GSM8K	0.093 [+0.074, +0.113]	0.609	+0.516	+0.504
Qwen2.5-1.5B, CommonsenseQA	0.657 [+0.606, +0.707]	0.633	−0.024	+0.039
Qwen2.5-1.5B, ARC-Challenge	0.677 [+0.629, +0.726]	0.661	−0.015	+0.028
Qwen2.5-14B, MATH-500 L3–5 (held out)	0.120 [+0.088, +0.154]	0.717	+0.597	+0.633

The readout was generated for every sampled-base arm, 11 in all and every row of Table 14, of which the Qwen2.5-14B arm was held out and the other 10 were fitted. Across the 10 fitted arms the readout effect and the chain-of-thought gain correlate at $r = 0.98$ with an ordinary-least-squares slope of 0.83; leaving any one arm out, r stays between 0.97 and 1.00. The correlation is close to an identity

by construction, since a forced readout at the first grid index has seen almost none of the chain and one at the last has seen almost all of it; that is the point, not a coincidence.

A registered out-of-sample test. A correlation over arms already run can be read as a description of those arms. To make it a claim about arms not yet run, the rule was registered first. `results/formal/prereg_14b.json`, committed before any 14B token was generated, records the fit on the 6 arms that existed then (the completion-API arm, Qwen2.5-1.5B on MATH-500, Qwen2.5-1.5B on GSM8K, Qwen2.5-1.5B on CommonsenseQA, Qwen2.5-3B and R1-Distill-1.5B), the rule $\Delta_{\text{readout}} = \text{slope} \cdot (\text{Acc} - \text{Acc}_{0\%}) + \text{intercept}$, and a ± 0.124 prediction band from that fit's residual spread (± 2.0 residual standard deviations of a 6-point fit, so the band is wide and the test it supports is correspondingly weak); the repository's history is the timestamp. That timestamp is self-recorded: the file was committed by us, not deposited with a registry, so a reader who does not trust the authors has the frozen fit and the rule but not independent evidence of when they were frozen. Qwen2.5-14B-Instruct was then generated on MATH-500 L3–5 under the settings of the 3B and 7B arms. Its chain is worth +0.597, so the rule predicted +0.520 [+0.396, +0.643]; the measured effect is +0.633, inside the band at +0.113 from its centre, near the upper edge rather than the middle, which is the honest description of a one-arm test with a band this wide. The sceptic's baseline does better still: the bare identity $\Delta_{\text{readout}} = \text{Acc} - \text{Acc}_{0\%}$, with no fitted slope or intercept at all, predicts +0.597 and lands +0.036 from the measurement, closer than the fitted rule. We report that rather than hide it, because it is the stronger form of the same claim: the intervention effect is not merely correlated with what the chain is worth, it is approximately equal to it, and the fitted slope adds nothing out of sample. The arm also carries the flat null at the largest token-exact scale in the paper (slope -0.0005 [-0.0056, +0.0046], equivalent to no change at 0.03), and it passes all five internal checks (Table 5). The interpretation is the one the flat null already forces and this makes concrete: a truncated-CoT accuracy curve traverses the distance between answering without the chain and answering with it, which is a fact about the task and the model's use of reasoning, not about the moment at which a particular trajectory settled. CommonsenseQA is the control: there the chain is worth -0.024 and the curve does not rise.

O THE SHAPE OF THE AGREEMENT CURVE

Proposition 2 fixes $G(0) = \mathbb{E}\|P_0\|^2$ and $G(t) \to 1$ for every model and task, so neither endpoint distinguishes anything; only the path between them can. For each trace we read $\hat{P}_t(Y)$ at the grid point nearest a given fraction of its own length, average over traces, and report $(\hat{G}(d) - \hat{G}(0))/(1 - \hat{G}(0))$: the fraction of the remaining distance to one that has been closed by depth d. This normalisation removes the intercept, which Appendix P shows accounts for most of the variation in the raw curve.

At half depth the closed fraction runs from 0.15 to 0.50 across the 15 exact-continuation arms of Table 4, so the shape is not common to all models and tasks. The flat null removes the accuracy curve as an instrument and Proposition 2 removes the endpoints of the agreement curve; the normalised shape survives both, is measurable on the corrected protocol, and is not common across models and tasks, so it is a quantity that can distinguish them, which is what a commitment claim needs and what neither of the two curves the literature plots can supply. A downstream claim built on it would read: *model A closes a larger fraction of its agreement distance by half depth than model B on the same task*, with both measured by exact continuation on an absolute grid and both passing the checks of Section 3. We do not make such a claim here, because doing it properly needs arms chosen to vary one thing at a time, which ours are not.

P EQUIVALENCE, POWER AND STRATIFIED FLATNESS

A null result needs a power statement. For each continuation arm we take the 95% cluster-bootstrap interval of the end-to-end change of $\hat{K}$ and report the smallest margin δ from $\{0.02, 0.03, 0.05, 0.10, 0.20\}$ for which the interval lies inside $(-\delta, +\delta)$. This is a descriptive statistic, not a test: δ is chosen after seeing the interval, so it is the interval's half-width rounded up to the grid, and it answers "at what margin would this arm pass a two-one-sided-test equivalence test at the 5% level" rather than asserting that it passes one at a margin fixed in advance. Read it as an upper bound on the end-to-end movement the data permit. All 11 sampled-base continuation arms,

which is every measurement continuation arm except the 4 greedy sub-arms, are equivalent to flat at 0.03–0.05; the 4 greedy sub-arms, with one trace per question, only at 0.10–0.20. A change of 0.028–0.036 would have been detected with 95% power on the largest arms, against published rises of 0.2–0.6.

A flat average could in principle hide rising and falling sub-populations. Stratifying by anything computed from the completed trace is not a valid check, because that is precisely the conditioning Proposition 4 says breaks the null: 17 of 25 strata defined by a trace's own length or by its question's realised accuracy have slopes excluding zero, in the direction the covariance term predicts. The valid checks condition on information available before generation, or on a disjoint half of the traces: over strata defined by the dataset's own difficulty label, by prompt length, and by a question's accuracy estimated on the other half of its traces, 32 of 33 have slope intervals containing zero. The exception is a stratum of R1-Distill-Qwen-1.5B, the one arm that fails the flat null overall.

Finally, Proposition 2 implies that the level of an agreement curve is set by its intercept. On our data the per-question Spearman correlation between $\hat{G}(0)$ and the area under the agreement curve is 0.82–0.96 under exact continuation and 0.21–0.82 under the forced readout, which is the regularity Lanham et al. (2023) report and use as a shortcut.

Q ENDPOINT AUDIT

The prefix is "*Let me solve this step by step. Weng earns $12 per hour. 50 minutes is 50/60 of an hour, so she*" sent as a trailing assistant message after the GSM8K question, at temperature 1.0 with 80 output tokens (1 500 for the two endpoints with hidden reasoning, which otherwise return no visible text), 24 samples per endpoint. A reply is classified as *continue* if its first non-space character is a lowercase letter (a grammatical continuation of "…so she"), as *acknowledge* if it is not a continuation but its first sentence refers to the prefix as prior text ("let's continue", "you're on the right track"), and as *ignore* otherwise. The rule is not about capitalisation: every reply classified as *continue* opens mid-sentence with a verb phrase ("worked", "earned", "earns"), and every reply classified as *ignore* or *acknowledge* opens a new answer instead—restating the problem, announcing a start, or giving the result outright. All raw replies are released so the classification can be re-derived, and Table 15 gives every endpoint's rates. An endpoint counts as *silently failing*, the 12 of Section 4.3, when it ignores the prefix in a majority of samples, which is 10 that never continue it plus the mixed endpoints whose ignore rate exceeds one half; the 3 that continue it in at least 95% of samples and the 2 that visibly acknowledge and restart are not silent failures, because a study using them would notice. 13 further endpoints, spanning 6 vendors beyond the 4 scored here, returned no sample and are absent rather than failed: 12 were unavailable to us (a model not served on our account, retired, rate-limited, or an empty reply), and 1 was reached and refused. `claude-opus-4-6` (Anthropic) answers a trailing assistant message with "*This model does not support assistant message prefill. The conversation must end with a user message*". That is the honest version of the failure Section 4.3 measures: the endpoint declines the substitute instead of silently discarding it, and a study using it would have been told rather than left with a plausible curve. All 13 replies are in the released audit file, so the classification here can be re-derived.

Table 15: The 18 audited endpoints: percentage of 24 samples that continue the prefix, acknowledge it and restart, or ignore it, and the class used in Section 4.3 (continues: $\geq 95\%$ continuation).

endpoint	continue %	acknowledge %	ignore %	class
doubao-seed-2-0-lite-260215-nothinking	100	0	0	continues
doubao-seed-2-0-mini-260215-nothinking	100	0	0	continues
gpt-3.5-turbo	100	0	0	continues
qwen-max	92	0	8	mixed
gpt-4o	21	0	79	mixed
deepseek-chat	4	0	96	mixed
qwen-turbo	0	100	0	acknowledges
qwen3-next-80b-a3b-instruct	0	100	0	acknowledges
deepseek-reasoner	0	0	100	ignores
gpt-4.1	0	0	100	ignores
gpt-4.1-mini	0	0	100	ignores
gpt-4.1-nano	0	0	100	ignores
gpt-4o-mini	0	0	100	ignores
gpt-5	0	0	100	ignores
gpt-5-mini	0	0	100	ignores
gpt-5-nano	0	0	100	ignores
o4-mini	0	0	100	ignores
qwen-plus	0	0	100	ignores

Continuing a prefill is necessary, not sufficient. The audit classifies whether an endpoint extends the prefix grammatically. That is a weaker property than reproducing the conditional law, because a prefill is re-tokenised inside the model's template even when it is honoured. Appendix T measures what that costs on a model where both arms are available: re-encoding the prefix leaves the correctness curve where it was and still moves the conservation statistic by -0.0124 [-0.0210, -0.0037] on the same cells. An endpoint's classification here is a screen; the within-arm checks are the test, and the one that catches a re-encoded prefix needs a continuation arm — which a closed chat endpoint does not provide. Two exploratory prefill files from before the completeness guard existed ship with the archive, described in `results/PILOT_NOTES.md`; no number in this paper reads them.

Access path, and what it can and cannot explain. Every closed-model endpoint in this paper—the completion-API arm of Table 3 and all 18 audited chat endpoints—was reached through one OpenAI-compatible endpoint aggregator rather than through each vendor's own API. The audit therefore measures prefill behaviour as it is observable through a single common access path, which is also the path a practitioner using such a service would meet. The obvious worry is that the aggregator, not the vendors, drops the trailing assistant message. Three observations bound it. First, behaviour through that one path is not uniform: 3 endpoints continue the prefix in every sample, 2 acknowledge it in every sample, 10 ignore it in every sample, and 3 sit in between at rates of 4%, 21%, 92%—five distinct behaviours that a path rewriting or dropping the message could not produce. Second, vendor-specific parameters and vendor-specific error bodies pass through unchanged: DeepSeek's documented `prefix` flag returns that vendor's own parameter error naming its beta base URL, in the vendor's own error schema, and Alibaba's `partial` flag is accepted by one Alibaba endpoint and rejected by two others, without changing any endpoint's behaviour (Section 4.3). Third, the arm the paper's measurements actually rest on, `gpt-3.5-turbo-instruct` through the completion interface, continues a text prefix exactly, which is the property the flat null and the conservation and terminal-agreement checks of Section 4.4 would detect the failure of. One possibility remains unexcluded: a path-specific effect that happens to reproduce vendor-shaped variation. A first-party replication of the audit would settle it and is the natural next step; we report the audit as a statement about endpoints as reached, not about vendor implementations.

R THE R1-DISTILL-1.5B DEVIATION BY GRID DEPTH

Table 16: Flat-null slope of the two reasoning arms when the grid is cut at increasing depth. "Alive" is the fraction of (trace, index) cells that are continuation samples rather than absorbed values.

arm	grid	alive	slope [95% CI]	
R1-Distill-1.5B	full, ≤ 4096	85%	-0.0102 [-0.0190, -0.0018]	excludes 0
	≤ 1024	96%	-0.0080 [-0.0160, +0.0001]	contains 0
	≤ 512	99%	-0.0044 [-0.0123, +0.0033]	contains 0
R1-Distill-7B	full, ≤ 4096	70%	+0.0012 [-0.0071, +0.0094]	contains 0
	≤ 1024	88%	+0.0066 [-0.0018, +0.0147]	contains 0
	≤ 512	98%	+0.0043 [-0.0031, +0.0114]	contains 0

S A READOUT ARTEFACT AT THE ABSORPTION BOUNDARY, IN OUR OWN ANALYSIS

An earlier version of this analysis reported a conservation gap on 6 continuation arms, of the same sign on every one and as large as -0.408. This appendix records what it was, because the answer is a protocol error of exactly the kind the paper is about, committed in our own readout, and because the statistic that found it is now one of the five checks.

Localising it. Eq. (6) makes the gap a sum over grid intervals, so a failing arm can be read rather than only reported. On every failing arm the gap sat almost entirely in the intervals where absorption enters, which are the intervals whose right-hand cell is δ_Y and in which the term reduces to $\hat{P}_s(Y) - \widehat{\|P_s\|^2}$. Evaluating that same quantity at every alive cell rather than at one interval per trace, which is the own-answer check of Section 3, put the whole deficit in the cells within eight tokens of the trace's own end, 0.1–2.8% of alive cells, where it reached -0.800 against a null of zero, while the remaining cells were null to a few thousandths.

What it was. A continuation record stores what the model wrote *after* the prefix. The answer the theory asks for is the answer of the whole trace, prefix and continuation together. Grading the continuation alone is exactly right wherever the continuation restates an answer, and exactly wrong in those last cells: there the base trace has already written its answer, the continuation adds a token or two and stops, and the grader reads a truncated number or nothing at all. Every one of the K draws of such a cell reads the same truncated string, so they agree with each other and disagree with the trace they continue, giving a large, consistent and entirely spurious $\hat{P}_s(Y) - \widehat{\|P_s\|^2} < 0$. The fix is to grade `base_text`$[: n_t]$ + continuation, where n_t is the character length of the decoded token prefix (`code/emit_prefix_chars.py`); `code/run_readout_artefact.py` runs the analysis both ways on all 15 arms and is what the numbers here come from.

Table 17: The conservation gap before and after grading the continuation together with its prefix. One row for each of the 6 continuation arms the old readout pushed out of the check, plus the control. "near" is the share of alive cells within eight tokens of the trace's end, the only cells at which the two gradings can differ. The emulated prefix-ignoring arm is the control: its "continuations" are fresh full traces, which always carry an answer of their own, so the correction leaves it where it was.

arm	near	gap, continuation only	gap, prefix + continuation
Qwen2.5-7B, MATH-500	1.0%	-0.097 [-0.166, -0.036]	+0.039 [-0.006, +0.084]
Qwen2.5-1.5B, GSM8K	1.1%	-0.143 [-0.208, -0.078]	+0.013 [-0.034, +0.061]
Qwen2.5-1.5B, CommonsenseQA	2.8%	-0.408 [-0.504, -0.314]	-0.069 [-0.137, -0.004]
Qwen2.5-14B, MATH-500	0.8%	-0.107 [-0.164, -0.048]	-0.011 [-0.055, +0.033]
Qwen2.5-1.5B, ARC-Challenge	2.0%	-0.340 [-0.420, -0.265]	-0.011 [-0.059, +0.036]
Qwen2.5-7B, prefix re-encoded	1.0%	-0.189 [-0.254, -0.128]	-0.052 [-0.093, -0.008]
prefix-ignoring (control)	0.4%	+0.460 [+0.394, +0.525]	+0.457 [+0.391, +0.524]

What is left. The correction returns 4 of the 6 arms to the check outright; own-answer calibration is null on every sampled-base continuation arm (largest $|\hat{D}|$ 0.021) and the conservation gap of every

such arm is within 0.070 of zero. Read against an exact martingale simulated at the same K, grid and trace lengths, whose gap over 40 seeds has mean -0.0008 (s.d. 0.029) and central 95% band [-0.068, +0.046] (`code/cons_reference_band.py`), those residuals are what this statistic does at this design. Corrected across arms, 1 of 15 measurement continuation arms is rejected, the greedy-base 7B arm at +0.159 [+0.067, +0.252], and both diagnostic arms are (the prefix-ignoring arm and the re-encoded-prefix arm). That is the behaviour a specification test should have: it rejects the two arms built to be mis-specified and the one arm whose base path is an arg-max trajectory rather than a draw, and nothing else.

Why it is worth a page. The gap was not noise, not finite K, and not the serving stack; Appendix T reports the conditions we varied looking for it. It was a readout convention that was wrong on one cell in a hundred, in the one place, the boundary where the trace ends, where the wrong cells all point the same way. The published protocols this paper criticises fail for structurally identical reasons: a readout that is applied where the quantity being estimated is already determined, and an average over cells selected by the very event being measured. Our own instance is reported here in full because a check that has never caught anything is not evidence that the pipeline is right.

T THE FIVE CHECKS, AND WHAT EACH ONE CATCHES

A check is only worth running if something it is meant to catch actually trips it, so each of the 4 errors below is realised as an arm on the same base traces. The table is not flattering to the apparatus and is not meant to be: 1 check catches all 4 errors, conservation, and it is the one that needs a continuation arm, which is exactly what the protocols under criticism lack. The re-encoded prefix is caught by conservation alone. The two checks a forced-readout study can run unaided, drift and terminal agreement, catch two of the four between them and neither catches the third or fourth. Own-answer calibration catches the error a careful reader is most likely to commit by accident, a greedy base pool, and does so with an interval per cell rather than one per trace, which is why it and not conservation localised the readout artefact of Appendix S. The argument for running all 5 is therefore not that each is indispensable but that the cheap ones are all a single-arm study has, and they do not cover the expensive one.

Table 18: The 5 checks of Section 3 against 4 protocol errors, each realised as an arm on the same base traces, every cell computed from the same per-arm results as Table 5. "forced readout" is the literature's intervention (Section 4.1); "prefix ignored" is the emulated readout of Appendix K; "prefix re-encoded" is exact continuation conditioned on the decoded and re-encoded prefix instead of its own token identifiers, which is what a completion endpoint or an honoured chat prefill does with a string; "greedy base path" continues an arg-max base trajectory, which is not a draw from P_t. The drift test rejects the prefix-ignoring arm, but not for the right reason: the only interval that survives correction is one in which absorbed cells enter the average, whose values come from the analyst's absorption convention and not from the readout, and below that region it finds nothing — as Eq. (4) predicts for pseudo-beliefs that are constant in expectation (Appendix K). It is terminal agreement that detects the readout itself.

check	needs a second arm?	forced readout	prefix ignored	prefix re-encoded	greedy base path
flat null (Corollary 1)	yes	**rejects**	passes	passes	passes
drift, $h \equiv 1$ (Eq. (4))	no	**rejects**	**rejects**	passes	passes
terminal agreement (Proposition 5)	no	passes	**rejects**	passes	passes
own answer (Proposition 2)	yes	–	passes	passes	**rejects**
conservation, $h = P_s$ (Eq. (5))	yes	**rejects**	**rejects**	**rejects**	**rejects**

Power tracks the size of the error. A specification test that rejected every forced readout equally hard, regardless of how much the readout changed, would be a formality. It does not. Drift rejects all 14 forced arms, but by amounts that track the intervention: across the 13 forced arms the number of intervals surviving Benjamini–Hochberg and the size of the arm's readout effect have rank correlation +0.92, from ARC-Challenge (effect 0.028, 1 interval) to R1-Distill-7B (effect 0.640, 10 intervals). On the tasks where the chain of thought is worth nothing there is almost no intervention to detect, and the test says so. The *power tracks the size* of the misspecification, which is what makes the rejections on the mathematical arms worth reading.

The re-encoded prefix. Exact continuation conditions on the prefix's own token identifiers. A closed model offers at best a string, and re-encoding it need not reproduce the token sequence the model emitted. We generated the whole Qwen2.5-7B arm both ways, with the same base traces, grid, decoding law and server and only the conditioning differing, and the two are indistinguishable on four of the five checks. The text arm's flat-null slope is +0.0009 [−0.0043, +0.0057] against +0.0010 [−0.0039, +0.0056]; it passes the drift test (0 intervals surviving Benjamini–Hochberg) where the token-exact arm passes it; terminal agreement passes on both (+0.250 against +0.260); and the two correctness curves agree to 0.014 at every index, differing significantly at 1 of 14, with an end-to-end paired difference of −0.013 [−0.025, −0.002] over 1 338 shared traces. The paired contrast is the more powerful comparison, since it removes the question and the trace, so it sees a difference the within-arm flat null cannot.

Conservation is the exception. The text arm's gap is −0.052 [−0.093, −0.008], against +0.039 for the token-exact arm, and the arm is rejected where the token-exact arm is not. The sharper comparison is paired: on the cells both arms share, re-encoding moves the level statistic $\hat{P}_t(Y) - \widehat{\|P_t\|^2}$ by −0.0124 [−0.0210, −0.0037], while the correctness curve does not move at all. Nothing else we varied does: on the same cells of the same base traces, turning automatic prefix caching off moves it −0.0074 [−0.0158, +0.0008] and quadrupling the draws per cell to $K = 16$ moves it −0.0087 [−0.0189, +0.0021], both consistent with zero, and a fresh replicate of the reported conditions gives +0.0041 [−0.0055, +0.0137] against −0.0022 [−0.0142, +0.0095] for the arm as originally generated. The conditioning event is what matters, not the serving stack. The practical reading is that a curve can pass every check a single-arm study can run and still be conditioned on the wrong event, and that the cheapest way to find out is the conservation check — which needs a continuation arm, and is therefore unavailable to precisely the studies that most need it.

U Does the correction change the ordering, or only the numbers?

A reader may grant the flat null and still suspect that the published protocol, wrong as it is, was ranking models correctly anyway — that the artefact is a monotone distortion. It is not.

Both instruments are put on one scale: "what fraction of the rise is finished by half depth". For the published one that is the forced-readout curve read at a fraction of each trace's realised length over the traces still generating — the three choices of Section 4.2 together — as a fraction of its own end-to-end rise at $f = 0.5$. For the corrected one it is the normalised agreement curve of Appendix O at relative depth 0.5. Both are unitless, both are read at half of the same realised length, and both are computed per question, so one cluster bootstrap over questions resamples them jointly.

Both instruments score 8 of the 10 arms, and on those the rank correlation is $\rho = +0.50$ [−0.15, +0.81] — positive, as one would expect of two measurements of the same traces, but with an interval that covers zero. They agree at the top (R1-Distill-7B is first on both) and not in the middle, which is where a comparison between models is actually made: on MATH-500 the 1.5B model is ranked 3 of 8 by the published instrument and 8 of 8 by the corrected one. We do not claim the corrected ordering is the true one — Proposition 2 constrains the endpoints of $\hat{G}$, not its shape, so the shape is a measurement and not a theorem. The claim is narrower and enough: the two instruments are not monotone transforms of each other, so correcting the protocol is not a matter of re-scaling published conclusions.

The remaining 2 arms have no published statistic at all. Where the chain of thought is worth nothing — CommonsenseQA and ARC-Challenge, Appendix M — the published curve has no rise to normalise, so the fraction completed at half depth is a ratio of noise. That is the flat null again, met in the place a practitioner would meet it: the instrument does not merely mislead, on the task where the chain carries no information it returns nothing at all, while the corrected curve is still defined and still ordered.

V The three errors, measured on their own

Survivorship is entirely the changing population — under absorption $\pi_t \hat{S}(t) + (1 - \pi_t)\hat{F}(t) = \hat{K}(t)$ identically — and plotting only the traces still generating turns the flat curve into one that falls by −0.064 to −0.464 across the 15 continuation arms, while on the traces alive at *both* ends the paired

drop is indistinguishable from zero on 13 of them (Appendix F). *Index leakage* is estimated directly by $\hat{K}(\rho_f) - \hat{K}(t_1)$, since $K(t_1) = \mathrm{Acc}$: at the pre-specified $f = 0.5$ it ranges from -0.012 to +0.015 with none of the five tests surviving Benjamini–Hochberg correction (Benjamini & Hochberg, 1995), and over all ten fractions it stays within 2.3 points (Appendix G). *Last-switch indices* need a null before they mean anything: read through each arm's own grid, a belief martingale whose answer is drawn only at the terminal step already puts 0.543–0.556 of the trace after the "decision" and a symmetric walk 0.574, so the published 35–70% (Datta et al., 2026) bears on commitment only through its distance from that null (KS 0.30–0.56; Appendix I).

W A SECOND MODEL FAMILY

Every other token-exact arm above is Qwen-derived (Qwen2.5-1.5B, 3B, 7B and 14B, and the two R1-Distill-Qwen models), and every one was served by one vLLM build on one machine. The flat null is a theorem and should not care; the internal checks, the level statistic and the readout–0%-CoT correspondence are empirical and might. Llama-3.1-8B-Instruct was therefore run on the same questions, grid, cap, R and K as the Qwen arms of Table 3.

Table 19: Llama-3.1-8B-Instruct on MATH-500 L3–5 beside the Qwen arms generated identically. Every column is the statistic defined where it is first used.

arm	Acc	slope of $\hat{K}$ [95% CI]	equiv.	$\Delta_{\mathrm{readout}}$	conservation
Llama-3.1-8B, MATH-500	0.359	-0.0037 [-0.0104, +0.0030]	0.05	+0.289	passes
Qwen2.5-7B, MATH-500	0.646	+0.0010 [-0.0039, +0.0056]	0.03	+0.450	passes
Qwen2.5-14B, MATH-500	0.717	-0.0005 [-0.0056, +0.0046]	0.03	+0.633	passes

What it shows. The flat null holds on the Llama arm (slope -0.0037 [-0.0104, +0.0030], equivalent to no change at 0.05), as it must, and the readout intervention is +0.289 [+0.245, +0.333], the same phenomenon at the same size as the Qwen arms of comparable accuracy. The conservation check passes on it and own-answer calibration does too (-0.0021 [-0.0108, +0.0066]), so nothing of Appendix S is specific to the Qwen family or to the way those arms were served. The drift check is the exception: 1 of its intervals survives Benjamini–Hochberg, the interval starting at $t = 128$ with a mean increment of -0.021, while the arm's slope and its end-to-end change both contain zero (-0.012 [-0.033, +0.010]). On a single arm the drift check therefore has a false-alarm rate we cannot claim is small, one of the two second-family and greedy-base arms that fail anything at all, which is why Section 3 runs five checks and not one.

The sharper result is the one this arm was not designed for. The regression of Appendix N says the readout intervention measures what the chain of thought is worth, and on a model from another family, another tokenizer and another instruction-tuning recipe the two land on top of each other: the chain is worth +0.284 and the readout effect is +0.289, a difference of +0.005. This arm is one of the 10 in the Appendix N fit, not a second held-out test, since the held-out arm there is the 14B one, so it is a new family entering the relationship rather than an independent check of it.