CONSERVED VERDICTS, DRIFTING COMMITMENTS: IDENTIFYING HARMFUL OVERTHINKING

Anonymous authors
Paper under double-blind review

ABSTRACT

When does additional inference compute improve an LLM's answer, and when does it erode an answer already available? We develop a conservation-based identification framework for this question. Under any fixed continuation policy, the exact forecast of terminal correctness is a martingale, including when continuation is harmful. Even its full trajectory cannot, by itself, identify the sign of stopping advantage. The missing quantity is the model's *current commitment*: correctness under a specified answer-if-stopped-now operator. Its gap from the conserved verdict equals both the negative of expected remaining drift and the conditional accuracy advantage of stopping. A second exact identity separates within-cohort change from termination-induced composition in budget curves. On MATH-500, eight fresh samples for each of 367 problems yield $100/2936$ early-correct/final-wrong 1.5B trajectories (3.4%; problem-clustered 95% CI $[2.4\%, 4.5\%]$). Across 1.5B/7B configurations, post-2048-token marginal declines of $8.3/4.4$ points combine positive within-cohort terms ($8.5/20.0$) with negative composition terms ($-16.8/-24.4$). Matched continuations from 12 previously wrong-ending prefixes and 12 controls reproduce all immediate correct answers but finish correct in 50.0% versus 97.9% of 192 fresh continuations: local expected loss at selected states, not a prevalence estimate. The verdict-probe control remains unresolved, and no tested stopping replay improves accuracy at lower token cost. The contribution is an identification foundation for evaluating inference-time compute, not a new early-exit policy or a claim that harmful overthinking is ubiquitous.

1 INTRODUCTION

Long chains of thought can improve reasoning (Wei et al., 2022; Snell et al., 2024; Muennighoff et al., 2025), but further computation does not guarantee a better answer. "Overthinking" describes several distinct observations: wasted computation on easy problems (Chen et al., 2024; Sui et al., 2025), accuracy reported to rise then fall with budget (Zhou et al., 2026), and within-chain abandonment of a correct answer (Wei et al., 2025).

The underlying development question is not whether a long trace ends correctly, but whether *continuing from its current state adds value*. These are different prediction targets. Confusing them can assign an outcome forecaster the wrong diagnostic task, attribute a changing evaluation population to failed reasoning, or suppress useful improvement in the name of stability. A theory of inference-time compute must therefore specify which observable identifies the benefit or harm of continuing.

Scope. A correct-to-wrong flip is a realized event; systematic execution drift is adverse predictable movement of answer-if-stopped-now correctness. A correct checkpoint is operator-relative, not evidence of answer recognition; prefix resampling tests expected loss (Sec. 5.3).

Overthinking is not identified by its natural measurements. Two tempting measurements do not isolate systematic erosion. A probe of final correctness (Zhang et al., 2025a) targets a value function whose exact increments have zero conditional mean under any fixed policy. Conversely, accuracy averaged over chains still running at each budget can decline as successful chains terminate, even without net degradation among survivors. Figure 1 separates these state-level and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

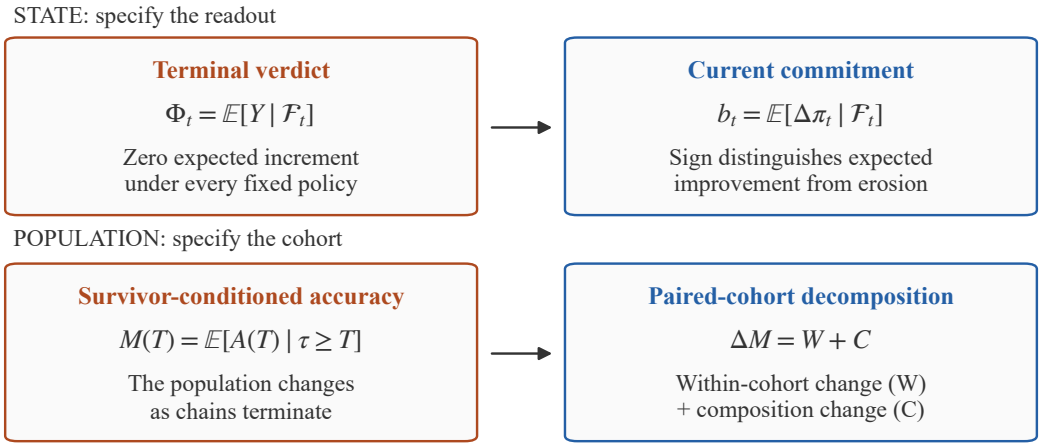


Figure 1: **Two levels of identification.** Verdict conservation does not imply commitment conservation: their gap, not verdict drift, identifies stopping value. At the population level, the cohort decomposition separates answer change from changing composition. Both distinctions precede questions of probe accuracy and stopping-policy performance.

population-level issues. The conserved verdict supplies a reference, not a law that answers remain correct: conservation holds for both helpful and harmful policies. Its importance is precisely this limitation. It forces a distinction between forecasting eventual success and measuring the value of more computation.

Logical route. We define expected answer change, establish the limits of outcome forecasts, and connect their gap from current quality to stopping value. A separate cohort identity handles population curves. Probe estimation is tested, not assumed (Sec. 5.6).

Contributions. **(1) A non-identification result and its resolution.** Even an exact terminal-outcome forecast does not determine whether stopping is better (Cor. 3). The commitment–verdict gap supplies the missing comparison and equals the negative of expected remaining drift (Prop. 2; Cor. 4). **(2) A population-level correction.** An exact decomposition separates answer evolution from endogenous termination (Prop. 5), identifying which part of a declining survivor curve concerns within-chain change. **(3) A behavioral audit of the distinction.** On MATH-500 (Hendrycks et al., 2021), DeepSeek-R1-Distill-Qwen 1.5B/7B (DeepSeek-AI et al., 2025) exhibit aggregate within-cohort improvement, while matched continuations establish local expected loss at selected correct prefixes. An unresolved probe control and unsuccessful stopping replays separate identification from reliable estimation and deployment.

Implications for reasoning models. *Evaluation* must separate answer change from sample composition; *process supervision* must distinguish outcome prediction from current answer quality; *compute allocation* must estimate continuation value, not infer it from confidence or length. Classical probability supplies the tools; our contribution identifies the targets and evidence connecting them. The drift objective illustrates a training implication, not a fine-tuning gain.

2 Commitment Dynamics and Drift Decomposition

A problem $q \sim \mathcal{D}$ is encoded as $h_0 = \mathrm{enc}_\theta(q)$, and reasoning proceeds by $h_{t+1} = F_\theta(h_t, \xi_t)$, where one step is one chain-of-thought segment (or one latent iteration) and ξ_t collects the sampling randomness. A fixed readout r exposes the *belief process* $\pi_t = r(h_t) \in [0,1]$: the probability that the model's *current* answer, elicited by a fixed stopping operator, is correct. Probability averages over that operator's sampling randomness; for a deterministic operator and a fully specified problem/prefix, correctness is binary. This is an evaluator-defined success probability, not necessarily a self-reported belief. Here h_t is the complete reasoning state, including the problem, prefix,

and budget; an experimental hidden-state tap can be lossy. Let $\mathcal{F}_t = \sigma(h_0, \ldots, h_t)$. A process (u_t) is *adapted* if u_t is $\mathcal{F}_t$-measurable; an increment process is *predictable* if its time-t term is $\mathcal{F}_t$-measurable, i.e. known one step ahead.

Assumption 1. *(A1) The sampling noise ξ_t is i.i.d. and independent of $\mathcal{F}_t$, so (h_t) is a time-homogeneous Markov chain with operator $(Pf)(h) = \mathbb{E}[f(h_{t+1}) \mid h_t = h]$; and terminal correctness Y is a function of the state in which the chain absorbs. (A2) The readout is a measurable function of h_t alone (lossy taps: Sec. 3.1); a critic receiving the step index augments the state to (h_t, t), for which all results hold.*

A changed answer does not reveal whether a step was harmful in expectation. To separate expected change from its realized fluctuation, condition on the information available before the step:

$$\Delta_t := \pi_{t+1} - \pi_t = b_t + \eta_t, \qquad b_t := \mathbb{E}[\Delta_t \mid \mathcal{F}_t], \quad \mathbb{E}[\eta_t \mid \mathcal{F}_t] = 0. \tag{1}$$

Here b_t measures expected answer improvement ($b_t > 0$) or erosion ($b_t < 0$), whereas η_t is a zero-conditional-mean residual. Doob uniqueness makes $B_t = \sum_{s<t} b_s$ the canonical predictable component of π_t (Thm. 7, App. D), once the readout and information history are fixed. This identifies a statistical quantity, not a cognitive mechanism: a lossy-tap residual can conceal predictable drift, and a generation–verification gap remains a possible explanation.

A *conserved commitment* has $b_t = 0$ throughout. This is not a general training goal: it forbids expected improvement as well as erosion, and even a constant wrong readout satisfies it. We instead need a reference that remains well-defined when computation helps or harms.

3 The Verdict Conservation Law

Conservation of terminal-outcome forecasts. The natural reference is the probability of the *same terminal event*, Y, as more of the reasoning history is revealed. An exact forecast already averages over future gains and losses under the continuation policy; its expected next revision must therefore be zero. This constrains a forecast of reasoning's outcome, not the quality of the answer currently available. The next result both establishes that distinction and connects the two quantities through expected remaining drift.

Proposition 2 (Conserved verdict)**.** *Let $Y \in \{0, 1\}$ be terminal correctness and define the* verdict $\Phi_t := \mathbb{E}[Y \mid \mathcal{F}_t]$.

(i) (Forecast coherence under any fixed policy.) *(Φ_t) is a martingale: $\mathbb{E}[\Phi_{t+1} \mid \mathcal{F}_t] = \Phi_t$ for every t, by the tower property. This needs* no *assumption on the dynamics, convergence or absorption, and holds for every fixed continuation policy,* including one that overthinks. *Its Doob decomposition therefore has $B_t \equiv 0$: the verdict has no predictable component to find.*

(ii) (Local change determines the global gap.) *Suppose the chain absorbs at τ with $\mathbb{E}[\tau] < \infty$ (automatic under a hard budget), is frozen thereafter ($\pi_s = \pi_\tau$, $b_s = 0$ for $s \geq \tau$), and commits to its realized outcome, $\pi_\tau = Y$. Then*

$$\pi_t - \Phi_t \;=\; -\,\mathbb{E}\left[\sum_{s \geq t} b_s \,\middle|\, \mathcal{F}_t\right], \tag{2}$$

so $\pi_t = \Phi_t$ for all t a.s. iff $b_t \equiv 0$, by uniqueness in Thm. 7: the commitment deviates from the conserved verdict by exactly the expected future accumulated drift.

Proof sketch. Part (i) is the tower property. For (ii), sum $\mathbb{E}[\pi_{s+1} - \pi_s \mid \mathcal{F}_t] = \mathbb{E}[b_s \mid \mathcal{F}_t]$ from $s = t$ to absorption and use $\pi_\tau = Y$. Boundedness and $\mathbb{E}[\tau] < \infty$ justify the limit; Doob uniqueness gives the equivalence (full proof in App. D).

Verdict-only non-identification. Zero verdict drift cannot certify harmless reasoning: it holds even when continuation erodes answers. More strongly, the verdict's values and entire trajectory need not supply the missing comparison, as the following construction shows.

Corollary 3 (Verdict-only non-identification)**.** *Under the conditions of Prop. 2(ii), the verdict trajectory alone need not identify the sign of $\pi_t - \Phi_t$ at an intermediate state. Two admissible commitment processes can have the same verdict trajectory and opposite conditional accuracy advantages of stopping.*

Proof by construction. Let $\tau = 2$, reveal $Y \sim \text{Bernoulli}(1/2)$ only at time 2, and fix $0 < \epsilon < 1/2$. Set $\pi_0^{\pm} = 1/2$, $\pi_1^{\pm} = 1/2 \pm \epsilon$, and $\pi_2^{\pm} = Y$. Both share $\Phi_0 = \Phi_1 = 1/2$, $\Phi_2 = Y$, but their time-1 stopping advantages are $\pm\epsilon$ and their remaining drifts are $\mp\epsilon$. A fixed randomized stopping operator can realize both commitments on two intermediate state types. □

At $\epsilon = 1/4$, both verdicts are 50%, but stopping yields 75% versus 25%. Current answer quality is missing, not forecast accuracy. This limits verdict-only diagnosis, not the usefulness of values for search.

3.1 The Commitment–Verdict Gap and the Poisson Equation

Harmonic reference and accumulated drift. The common terminal boundary in Prop. 2(ii) makes the gap equal to accumulated *remaining* expected change. In Markov form, the same relationship becomes a Poisson equation. Under Assumption 1, the exact value function $\Phi(h) = \Pr(Y = 1 \mid h)$ satisfies $P\Phi = \Phi$: it is harmonic for the fixed continuation policy. Therefore

$$b = (P - I)\pi = (P - I)(\pi - \Phi). \tag{3}$$

In a finite absorbing chain, Φ is the unique bounded harmonic extension of terminal correctness and $\pi - \Phi = -(I - Q)^{-1}b$, with Q the transient-state matrix (App. E). Thus the gap weights signed local drift by expected future state visits; positive and negative changes can offset. In particular, $|\pi(h) - \Phi(h)| \le \|b\|_\infty \mathbb{E}_h[\tau]$. This explains why one adverse step need not make all remaining computation harmful. Drift belongs to the readout–dynamics pair, not hidden-state motion alone.

Probe estimation and observation loss. Identification specifies the exact quantity; estimation must recover it. For a lossy tap $z_t = f(h_t)$, a critic observes only $m_t = \mathbb{E}[\Delta_t \mid z_t, t]$; conditional Jensen gives $\mathbb{E}[m_t^2] \le \mathbb{E}[b_t^2]$. This population bound concerns the *same* readout, not an arbitrary learned substitute. If $\hat{\pi}_t = \pi_t + e_t$, its full-filtration drift differs by $\mathbb{E}[e_{t+1} \mid \mathcal{F}_t] - e_t$, at most 2ϵ when $|e_t| \le \epsilon$ uniformly (Prop. 14). AUC and binned calibration error do not establish that bound.

The commitment probe predicts current forced-answer correctness; the verdict probe predicts terminal correctness. An exact verdict conditioned on a nested observation history is a martingale, even without full-state sufficiency. Current-state taps need not retain that history: even $v_t = \mathbb{E}[Y \mid z_t, t]$ can have predictable increments (App. I). Thus calibration alone does not establish the verdict null. Our control remains unresolved (Sec. 5.6); matched behavioral continuations provide a probe-independent check on loss.

4 Decision and Training Implications

The gap has decision-theoretic meaning because its two terms score the two actions at the same state: answer now, or follow the fixed continuation policy. The next identity turns that comparison into a criterion for accuracy-improving stopping.

Corollary 4 (Stopping advantage is remaining adverse drift)**.** *Under Prop. 2(ii), let $S \le \tau$ be a stopping time and $C_S \in \{0, 1\}$ the correctness of the stopping operator's answer, with $\mathbb{E}[C_S \mid \mathcal{F}_S] = \pi_S$. Then*

$$\mathbb{E}[C_S - Y] = \mathbb{E}[\pi_S - \Phi_S] = -\,\mathbb{E}\left[\sum_{t=S}^{\tau-1} b_t\right]. \tag{4}$$

In particular, if $b_t \ge 0$ a.s. before termination, no such stopping rule improves expected accuracy over the fixed continuation policy.

Proof. Apply the tower property to C_S and Y, then Eq. (2) at S; integrability follows from $\mathbb{E}[\tau] < \infty$. □

A stopping rule improves accuracy exactly when its selected states have a positive mean commitment–verdict gap. This is a criterion, not an observable oracle. Positive *average* drift across all states does not preclude selective gains. The decision must use only information available at S, under a shared continuation protocol.

The value of further compute. If continuing incurs nonnegative integrable cost K_S and $\lambda \geq 0$ is its price per unit, its conditional utility relative to stopping is

$$\underbrace{\Phi_S - \pi_S}_{\text{expected accuracy gain}} \quad -\lambda\, \mathbb{E}[K_S \mid \mathcal{F}_S] = \mathbb{E}\left[\sum_{t=S}^{\tau-1} b_t - \lambda K_S \,\middle|\, \mathcal{F}_S\right].$$

Thus an efficiency-motivated exit can be optimal even when expected accuracy improves: it is not evidence of harmful reasoning. Conversely, a negative one-step drift need not imply negative remaining value. These are policy-relative comparisons, not an implementable oracle; changing the continuation policy changes its verdict and requires a new evaluation.

Training implication: stability is not improvement. The decomposition also distinguishes which changes a regularizer suppresses. A quadratic increment penalty charges both predictable change and innovation: $\mathbb{E}[\Delta_t^2] = \mathbb{E}[b_t^2] + \mathbb{E}[\eta_t^2]$ (Prop. 15). Because $m_t = \mathbb{E}[\Delta_t \mid z_t, t]$ is an L^2 projection, $\mathbb{E}[m_t^2] = \max_g\{2\,\mathbb{E}[g(z_t,t)\Delta_t] - \mathbb{E}[g(z_t,t)^2]\}$ with $g^\star = m_t$, and penalizing this variational value—requiring no new labels given the readout, blind to innovation variance—vanishes iff the chain is *state-calibrated* ($m_t \equiv 0$). This symmetric objective penalizes beneficial predictable improvement too. For erosion alone, a sign-constrained variant targets $\mathbb{E}[(m_t^-)^2] = \sup_{g\leq 0}\mathbb{E}[2g\Delta_t - g^2]$, where $m_t^- = \max\{-m_t, 0\}$ (App. F); we derive, but do not experimentally evaluate, this variant. Suppressing movement, drift, and erosion are different objectives (Apps. F–G); only synthetic validation is provided (Sec. 5.1).

The critic as a sensor. Regress frozen-model probe increments on (z_t, t) and accumulate $\hat{B}_t = \sum_{s<t} g_\phi(z_s, s)$. With a sufficient tap and uniform critic error δ, compensating the exact readout by $\tilde{\pi}_t = \pi_t - \hat{B}_t$ gives $|\,\mathbb{E}[\tilde{\pi}_\tau - \tilde{\pi}_0]| \leq \delta\,\mathbb{E}[\tau]$ (Prop. 11, App. D). This is a readout-error bound, not a guarantee of improved answers or of accurate learned probes. Our diagnostic *drift alarm* stops when $-\hat{B}_t > \kappa$, using the last checkpoint answer; κ is pre-registered as the 90th percentile of $\max_t(-\hat{B}_t)$ over stay-correct critic-split chains. Accumulated past drift is a heuristic signal, not the remaining-drift oracle above.

Minimal audit recipe. (i) Elicit answers at fixed intermediate budgets. (ii) Measure correct-to-wrong prevalence with repeated samples within problems. (iii) Separate within-cohort and composition changes. (iv) Test a sensor or stopping rule where behavioral checks show expected loss at candidate states; aggregate deterioration is not required. The first three steps assess measurement; the fourth assesses control (App. G).

5 Experiments

Scope of the empirical evaluation. The identification results are theorems; the experiments test whether their audit is implementable from residual-stream taps and what it finds. The pre-registered estimands are predictable commitment drift, correct-to-wrong prevalence, and stopping-policy utility. The survivor decomposition, matched-prefix follow-up, and refit diagnostic are post hoc. Table 1 gives the primary outcomes; the full matrix is in App. B.

5.1 Experiment 0: apparatus calibration with known injected drift

A binary Gaussian-evidence process supplies known drift: its unperturbed posterior on a fixed hypothesis is a martingale, while a constant log-odds lean injects probability-scale drift. With evidence separation 0.35, noise SD 1, and 60 steps, the critic recovers it (slope 1.04, $R^2 = 0.82$ at $\varepsilon = 0.05$), reports no false drift at zero, and has matched-sample detection floor 0.05. Under decaying evidence, accuracy loses 15.8 points after peaking at $T^\dagger = 14$; compensation restores it to 85.3%. The variance-blind penalty preserves gain and accuracy over the tested λ values, whereas $\mathbb{E}[\Delta\pi^2]$ collapses both. This calibrates the apparatus, not real-model decision accuracy; the 19-configuration sweep and late-onset test are in App. H.

Table 1: Main audit results. Brackets are 95% CIs. Accuracy changes are percentage points; decomposition terms are summed over the four 512-token intervals from 2048 to 4096. The configurations differ in both precision and difficulty.

	R1-Distill-Qwen 1.5B	**R1-Distill-Qwen 7B**
Precision / MATH levels	FP16 / 3–5	NF4 / 4–5
Primary problems (one chain each)	367	262
Early-correct/final-wrong	$10/367$; $[1.5\%, 4.9\%]$	$9/262$; $[1.8\%, 6.4\%]$
Strict label (two matching checkpoints)	$6/367$	$3/262$
Survivor-conditioned change	-8.3	-4.4
Within-cohort component	$+8.5$ $[+2.1, +14.5]$	$+20.0$ $[+12.6, +27.9]$
Composition component	-16.8 $[-22.3, -12.0]$	-24.4 $[-31.9, -17.6]$
Commitment score $\widehat{V}/\widehat{\mathrm{Var}}(\Delta)$	0.284 $[0.263, 0.305]$	0.221 $[0.200, 0.245]$
Verdict score (control unresolved)	0.249	0.244
Commitment minus verdict score	0.036 $[+0.007, +0.063]$	-0.023 $[-0.050, +0.007]$

5.2 Real-model setup

We run DeepSeek-R1-Distill-Qwen-1.5B in FP16 on MATH-500 (the forcing operator closes an explicit thinking block, so the audit targets reasoning-trained models with one) and, after the pre-registered prevalence gate failed (Gate 1(a): at least 30 overthink-wrong chains on levels 3–5, else escalate to 7B on levels 4–5), DeepSeek-R1-Distill-Qwen-7B in 4-bit NF4 on levels 4–5. The configurations differ in precision and difficulty range, so this is not a controlled scaling comparison. One sampled chain per problem (temperature 0.6, top-p 0.95) runs to 4096 tokens; every 512 tokens, a separate branch closes the thinking block and requests a boxed final answer, decoded greedily for at most 48 tokens (Muennighoff et al., 2025). Branches do not modify the continuing chain. At termination or the cap, we use its boxed answer after the thinking block if present; otherwise we apply the same forcing rule. Normalized string matching and symbolic verification determine correctness; missing or unparseable answers count as incorrect. Relative to this operator, overthink-wrong means correct at a checkpoint ≤ 2048 (strictly before the end) and wrong finally; stay-correct means correct there and finally. The strict variant requires two consecutive early correct checkpoints with the same normalized answer. Wrong-ending chains without an early correct checkpoint are outside this early-erosion label; they may still have a later transient correct answer. Forced-answer correctness is the probe-independent behavioral signal, and the calibrated commitment probe that tracks it densely reaches an area under the ROC curve (AUC) of 0.68/0.75 (expected calibration error, ECE, 0.074/0.076); the critic and the verdict probe are introduced where they are used (Secs. 5.3, 5.6). Stratified problem-level probe/critic/eval splits are disjoint (40/30/30). Logistic probes use layer 21 checkpoint states and are temperature-calibrated within the probe split; the two-layer, 256-unit MLP critic learns paragraph-level probe increments from state and token position. Unless identified as Wilson intervals, CIs use 2000 bootstrap resamples of whole chains (whole problems for repeated samples), conditional on fitted probes and critics. "Pre-registered" denotes an internal dated analysis plan, not an external registry. Full hardware and implementation details are in App. H.

5.3 Q1: Do realized flips correspond to expected loss?

Instruments: forced-answer correctness at each checkpoint, and an MLP critic fit to commitment increments on uncorrected rollouts to give drift a sign.

Prevalence: Gate 1(a) fails at both scales. At 1.5B, $10/367$ chains are overthink-wrong (6 strict); the 7B escalation gives $9/262$ (3 strict): approximately 3% on the 512-token grid (Wilson 95% CIs $[1.5\%, 4.9\%]$/$[1.8\%, 6.4\%]$). Eight fresh samples per problem agree (3.4%, problem-clustered 95% CI $[2.4\%, 4.5\%]$), but only 50% of initially overthink-wrong problems repeat the label in any of eight, so one trajectory is not a stable problem label. The multiplicity- corrected per-problem budget screen is underpowered (App. H). Behavioral checks on Qwen3-1.7B, R1-distilled Llama-3.1 8B, and Phi-4-mini give 3.0% ($[1.7\%, 5.3\%]$), 2.2% ($[1.1\%, 4.2\%]$), and 2.2% ($[1.1\%, 4.2\%]$), respectively.

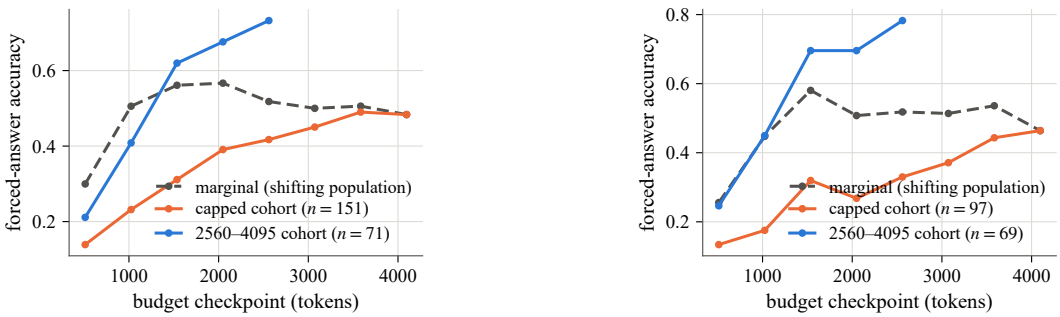


Figure 2: The survivorship artifact (Prop. 5) at 1.5B, levels 3–5 (**left**) and 7B, levels 4–5 (**right**). The *survivor-conditioned* curve (gray) declines after 1.5–2k tokens; fixed-cohort curves are flat-to-rising, with final-step deltas $-0.7\,[-4.0, +2.6]$ ($n = 151$) and $+2.1\,[-1.0, +6.2]$ ($n = 97$) points.

Predictable probe increments, with positive mean predicted drift. Here $\Delta = \Delta\hat{\pi}_t$ and $m = \mathbb{E}[\Delta \mid z_t, t]$ refer to the *probe* increments. For any fixed critic g, the population value $V(g) = \mathbb{E}[2g\Delta - g^2] = \mathbb{E}[m^2] - \mathbb{E}[(g-m)^2] \le \mathbb{E}[m^2]$ is a lower bound on state-conditional drift energy. Its held-out sample estimate $\widehat{V}$ is not itself a guaranteed finite-sample bound. We report the normalized score $\widehat{V}/\widehat{\mathrm{Var}}(\Delta)$: $0.284\ [0.263, 0.305]$ at 1.5B and $0.221\ [0.200, 0.245]$ at 7B, with chain-bootstrap intervals conditional on the fitted probes and critics. This variance normalization is not a bounded fraction of increment energy. Mean predicted increments are positive in all four label groups, with stay-correct CIs excluding zero ($[+0.0032, +0.0079]/[+0.0111, +0.0202]$). The critic also predicts positive mean drift for overthink-wrong chains and cannot resolve their rare erosion. Evidence for aggregate net improvement comes from the behavioral within-cohort term (Sec. 5.4); it does not exclude adverse drift in particular states. Randomized-refit diagnostics and their inferential limitations are in App. H.1.

Matched continuations establish loss at selected correct states. In a post-hoc follow-up, we revisit all 12 wrong-ending 1.5B chains with a correct checkpoint at or before 2048 (all five MATH levels), using the first such checkpoint. Each is paired with a unique right-ending control, correct at the same checkpoint and matched exactly on difficulty level. Selection and eight continuations per prefix were fixed before new outcomes were generated. Every immediate greedy answer reproduces its archived correct label. Continuations retain the original prefix tokens, sample at temperature 0.6/top-p 0.95, and receive $4096 - t$ tokens at checkpoint t. At EOS or the cap we extract the post-think boxed answer, or force one greedily for at most 48 tokens, as in the primary audit.

Of 96 continuations per arm, 48 cases and 94 controls finish correct: 50.0% versus 97.9%. Thus mean immediate-minus-continuation gaps are 50.0 and 2.1 percentage points; the paired contrast is 47.9 (95% paired-prefix bootstrap interval $[27.1, 69.8]$, 5000 resamples). All continuations produce a scored answer, with no correctness disagreement under strict balanced-box parsing. Under the matched protocol, the case gap estimates adverse expected remaining drift in (2), without relying on a probe.

This is a selected-state result, not an updated prevalence estimate or a deployable stopping rule: selection uses archived final outcomes, matching does not remove all problem differences, and eight draws leave wide per-prefix intervals. It neither identifies a cognitive mechanism nor conflicts with aggregate improvement. App. J gives all pairs, and uncertainty; App. H.1 reports the earlier cross-protocol diagnostic.

5.4 Q2: Does the declining budget curve reflect within-chain deterioration?

Instrument: the same forced-answer curve under two conditioning sets—survivors at each budget versus a fixed cohort—plus a forcing-free natural-length check.

Paired-cohort comparison and population selection. The state-level results compare actions at the same prefix. A population curve can violate this comparability by changing the evaluated chains.

Table 2: Held-out stopping replay: accuracy (%) and mean charged tokens, on 110/80 evaluation chains at 1.5B/7B with only two early-correct/final-wrong chains each. Small differences do not establish a policy ranking; App. H.1 gives the full 48-configuration gauntlet.

	R1-Distill-Qwen 1.5B		R1-Distill-Qwen 7B	
Policy	Accuracy	Tokens	Accuracy	Tokens
Full budget	77.3	2808	77.5	3015
Drift alarm	74.5	2624	73.8	2767
Checkpoint vote	71.8	3087	65.0	3315
Answer agreement	65.5	1625	61.3	1535

In Fig. 2, $M(T) = \mathbb{E}[A(T) \mid \tau \geq T]$ averages forced-answer correctness only over chains reaching T. Its decline after 1.5–2k tokens therefore mixes two distinct changes:

Proposition 5 (Survivorship decomposition)**.** *For budgets $T < T'$,*

$$M(T') - M(T) = \underbrace{\mathbb{E}[A(T') - A(T) \mid \tau \geq T']}_{\text{within-cohort}} + \underbrace{\mathbb{E}[A(T) \mid \tau \geq T'] - \mathbb{E}[A(T) \mid \tau \geq T]}_{\text{composition}}$$

(add and subtract $\mathbb{E}[A(T) \mid \tau \geq T']$). The first term is wrong-to-correct probability minus correct-to-wrong probability; the second is induced by termination. $M(T)$ is survivor-conditioned, not a deployment curve that carries a terminated chain's answer forward.

The identity is an accounting result, not a claim that every decline is an artifact: only the composition term is removed by holding membership fixed. It is orthogonal to the readout problem above; both must be controlled. The within-cohort term over $2048 \rightarrow 4096$, summed along the 512-token grid (App. H.1), is *positive* ($+8.5$ points $[+2.1, +14.5]$ at 1.5B; $+20.0$ $[+12.6, +27.9]$ at 7B): still-reasoning chains keep improving, and composition (-16.8/-24.4 points) more than accounts for the declines (-8.3/-4.4): aggregate net improvement over this grid, without excluding local erosion. Each interval uses its own later-survivor cohort; the sum is not the endpoint change for a single fixed cohort. The latter is shown separately in Fig. 2. Majority-vote curves inherit the same structure (within-cohort $+12.0$ points $[+4.6, +20.2]$; App. H.1). Qwen3-1.7B, Llama and Phi reproduce it ($+7.6$/$+16.6$/$+9.1$ within, -10.2/-25.9/-16.6 composition, in that order), while a forcing-free natural-length analysis finds the apparent decline dominated by censoring: aggregate slopes of -9.1/-8.6 points per 512-token bin fall to -2.4 $[-4.5, -0.6]$ and -1.1 $[-2.8, +0.3]$ on naturally terminated chains, the long bins being mostly capped chains (41%/37%).

5.5 Q3: DOES DRIFT-AWARE STOPPING HELP?

We fix thresholds on the critic split and replay policies using only information available at each decision, on held-out evaluation chains ($n = 110/80$ at 1.5B/7B). Table 2 compares the drift alarm, two-consecutive-answer agreement, checkpoint voting, and full budget. Charged tokens are generated length plus 48 per forcing branch, with no terminal forcing charged to the archived full-budget baseline; prefill and critic costs are excluded. This is not a wall-clock or FLOP comparison.

None of the implementable configurations ($44/48$ at 1.5B; $21/28$ at 7B), or alarm thresholds from the 50th to 99th healthy-chain percentile, improves accuracy at lower charged token cost. Token-saving rules lose -0.9 to -45.5 points at 1.5B and -1.3 to -47.5 at 7B; only $4/21$ and $3/16$ intervals reach zero. The alarm fires on $13\%/14\%$ of stay-correct chains and catches none/one of the two evaluation overthink-wrong chains. App. H.1 documents native-method deviations and the full replay matrix. These results assess our cached-checkpoint protocol, not the cited methods under their native execution.

The negative result distinguishes a quantity's theoretical relevance from our ability to estimate and act on it. It establishes no general impossibility of beneficial early exit.

5.6 VALIDITY: THE UNRESOLVED VERDICT CONTROL, AND ROBUSTNESS

The tower-property control remains unresolved. Score ordering changes across scales and taps (Table 1). Ten whole-problem bootstrap refits per model, using disjoint fitting, calibration and vali-

dation problems and a fixed evaluation split, give commitment above verdict in only $4/10$ and $1/10$ runs. This is tuning sensitivity, not a confidence interval (App. K). Verdict AUCs ($0.747/0.735$) also miss the prespecified 0.75 bar. An exact verdict on nonnested observations can drift (App. I); adding a cumulative history summary leaves negative gaps ($-0.067/-0.081$), but does not guarantee nested information. Neither test separates observation loss from estimation error (App. H.1).

Sensitivity of the behavioral findings. Absent the forcing string, 69% of 160 truncated prefixes emit no answer at all—the reason an operator is needed; where an unforced continuation does answer, it agrees with the forced answer on correctness in 85.7% of cases, and greedy forcing exceeds sampled by $+0.7$ points ($[-0.3, +1.8]$). The overthink-wrong count is grid-dependent (5/5/7 at 1024/512/256 tokens, matched $n = 240$), so $\approx 3\%$ is specific to the 512-token grid. Prevalence stays at 1–4% under every final-outcome definition and at temperature 0.9, the within-cohort term stays positive, and a linear critic retains the drift signal; on AIME-2024 (120 chains) no chain is overthink-wrong, though only 12 are ever correct early (App. H). Table 3 and App. H.1 give each check with its interval.

6 Related Work

Martingale views. Li et al. (2026a) use Doob decomposition and stopping theory for path selection; Kim et al. (2026) study coherence under answer resampling; Farmer et al. (2026) characterize elicited probability transformations. He et al. (2025) test self-reported belief revisions against a martingale rationality criterion. We add the distinction between a terminal-outcome forecast, conserved for every fixed policy, and an operational success probability, which may legitimately improve, and identify the comparison needed to value continued computation. The advance is in the estimands and their identification, not a new martingale theorem.

Overthinking and early exit. Wasted computation (Chen et al., 2024; Sui et al., 2025), budget-dependent accuracy (Zhou et al., 2026), answer abandonment (Wei et al., 2025), and agentic over-reasoning (Cuadron et al., 2025) are distinct targets. Our selection warning applies to survivor-conditioned curves only: it neither overturns controlled evidence of harm nor covers carried-forward fixed-cohort curves. Early-exit work spans convergence, redundancy detection, attention probes, cost-aware stopping and adaptive sampling (Zhang et al., 2025a; Liu & Wang, 2025; Mao et al., 2025; Kadavath et al., 2022; Sun et al., 2026; Min et al., 2026; Li et al., 2026b; Dong et al., 2026; Zabounidis et al., 2026; Aggarwal et al., 2023; Huang et al., 2025). Our audit separates evidence for erosion from any one stopping rule's utility.

7 Implications, Limitations, and Conclusion

Our contribution connects state-level reasoning dynamics to compute evaluation: verdict conservation limits what outcome forecasts diagnose, the commitment–verdict gap identifies stopping value, and the survivor decomposition separates answer evolution from population selection. Aggregate improvement and selected-state harm can therefore coexist.

Beyond single-chain reasoning. A step can be a reasoning segment, a self-revision round (Madaan et al., 2023), or a multi-agent exchange (Du et al., 2023). Under a fixed overall protocol, a nested joint history and a compatible stopping readout, the identities distinguish forecasting eventual success from valuing another round. For collaboration, the target is the system's deliverable answer, not agreement among agents. Recursive system improvement additionally requires explicit update rules and a fixed evaluation target; conservation of that target's forecast is not conservation of model capability. App. M formalizes these connections and testable audits. They are theoretical extensions, not experiments reported here.

Limitations. The probe control remains unresolved; stopping replays show no accuracy gain at lower cost. Guarantees are readout- and policy-relative. Probes and replays cover two R1-Distill-Qwen models; behavioral checks extend to Qwen3, Llama and Phi, mainly on MATH-500. The 4096-token cap binds often (41%/37%); larger-scale generalization and real-model drift-penalty fine-tuning remain untested.

Reproducibility statement

All experiments' models and dataset are public. Seeds (0 for the primary audits; the resampling and ablation studies draw from disjoint seed streams recorded in the released code), prompts and forcing strings, splits, hyperparameters, and the pre-committed gates are in App. H. The matched follow-up and refit variant have separate frozen settings, raw outputs, and checks (Apps. J–K); their seed streams start at 20260907. Quantitative tables are checked against result files. Internal dated plans are distinguished from external registration. The new follow-up code and results accompany the primary audit artifacts.

AI Use Statement

An AI coding assistant was used throughout this project: for drafting and editing prose, writing and debugging experiment and analysis code, scaffolding statistical analyses, and assisting with mathematical reformulation and proof checking. The authors take responsibility for the final text, theoretical claims, proofs, code, analyses, and reported results, including any errors.

References

Aradhye Agarwal, Ayan Sengupta, and Tanmoy Chakraborty. First finish search: Efficient test-time scaling in large language models. *arXiv preprint arXiv:2505.18149*, 2025.

Pranjal Aggarwal, Aman Madaan, Yiming Yang, and Mausam. Let's sample step by step: Adaptive-consistency for efficient reasoning and coding with LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 12375–12396, 2023.

Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. Do NOT think that much for 2+3=? on the overthinking of o1-like LLMs. *arXiv preprint arXiv:2412.21187*, 2024.

Alejandro Cuadron, Dacheng Li, Wenjie Ma, Xingyao Wang, Yichuan Wang, Siyuan Zhuang, Shu Liu, Luis Gaspar Schroeder, Tian Xia, Huanzhi Mao, Nicholas Thumiger, Aditya Desai, Ion Stoica, Ana Klimovic, Graham Neubig, and Joseph E. Gonzalez. The danger of overthinking: Examining the reasoning-action dilemma in agentic tasks. *arXiv preprint arXiv:2502.08235*, 2025.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Nishanth Dikkala, Greg Lewis, Lester Mackey, and Vasilis Syrgkanis. Minimax estimation of conditional moment models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Zhe Dong, Fang Qin, and Manish Shah. When does learning to stop help? a cost-aware study of early exits in reasoning models. *arXiv preprint arXiv:2606.30852*, 2026.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*, 2023. URL https://arxiv.org/abs/2305.14325.

Mike Farmer, Abhinav Kochar, and Yugyung Lee. Empirical characterization of inference-time elicited probability transformations in large language models. *arXiv preprint arXiv:2603.19262*, 2026.

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.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

Zhonghao He, Tianyi Qiu, Hirokazu Shirado, and Maarten Sap. Martingale score: An unsupervised metric for Bayesian rationality in LLM reasoning. In *Advances in Neural Information Processing Systems (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 *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021.

Chengsong Huang, Langlin Huang, Jixuan Leng, Jiacheng Liu, and Jiaxin Huang. Efficient test-time scaling via self-calibration. *arXiv preprint arXiv:2503.00031*, 2025.

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

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

John G. Kemeny and J. Laurie Snell. *Finite Markov Chains*. Undergraduate Texts in Mathematics. Springer-Verlag, New York, 1976. Reprint of the 1960 Van Nostrand edition, with a new appendix.

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.

Greg Lewis and Vasilis Syrgkanis. Adversarial generalized method of moments. *arXiv preprint arXiv:1803.07164*, 2018.

Huayu Li, ZhengXiao He, Siyuan Tian, Jinghao Wen, and Ao Li. Martingale foresight sampling: A principled approach to inference-time LLM decoding. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3522–3533, 2026a. doi: 10.18653/v1/2026.eacl-long.162. URL https://aclanthology.org/2026.eacl-long.162/.

Jiakai Li, Ke Qin, Rongzheng Wang, Yizhuo Ma, Qizhi Chen, Muquan Li, and Shuang Liang. Stop when further reasoning won't help: Attention-state adaptive generation in reasoning models. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026b.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Yiwei Li, Peiwen Yuan, Shaoxiong Feng, Boyuan Pan, Xinglin Wang, Bin Sun, Heda Wang, and Kan Li. Escape sky-high cost: Early-stopping self-consistency for multi-step reasoning. In *International Conference on Learning Representations (ICLR)*, 2024.

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. In *International Conference on Learning Representations (ICLR)*, 2024.

Xin Liu and Lu Wang. Answer convergence as a signal for early stopping in reasoning. *arXiv preprint arXiv:2506.02536*, 2025.

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

Minjia Mao, Bowen Yin, Yu Zhu, and Xiao Fang. Early stopping chain-of-thoughts in large language models. *arXiv preprint arXiv:2509.14004*, 2025.

Dehai Min, Giovanni Vaccarino, Huiyi Chen, Yongliang Wu, Gal Yona, and Lu Cheng. Stop when reasoning converges: Semantic-preserving early exit for reasoning models. *arXiv preprint arXiv:2605.17672*, 2026.

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.

Jürgen Schmidhuber. Learning factorial codes by predictability minimization. *Neural Computation*, 4(6):863–879, 1992. doi: 10.1162/neco.1992.4.6.863.

Adam S. Shai, Sarah E. Marzen, Lucas Teixeira, Alexander Gietelink Oldenziel, and Paul M. Riechers. Transformers represent belief state geometry in their residual stream. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, Na Zou, Hanjie Chen, and Xia Hu. Stop overthinking: A survey on efficient reasoning for large language models. *Transactions on Machine Learning Research (TMLR)*, 2025.

Renliang Sun, Wei Cheng, Dawei Li, Haifeng Chen, and Wei Wang. Stop when enough: Adaptive early-stopping for chain-of-thought reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 27250–27268, 2026.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

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

Zihao Wei, Liang Pang, Jiahao Liu, Wenjie Shi, Jingcheng Deng, Shicheng Xu, Zenghao Duan, Fei Sun, Huawei Shen, and Xueqi Cheng. The evolution of thought: Tracking LLM overthinking via reasoning dynamics analysis. *arXiv preprint arXiv:2508.17627*, 2025.

Haoran Xu, Baolin Peng, Hany Awadalla, Dongdong Chen, Yen-Chun Chen, Mei Gao, Young Jin Kim, Yunsheng Li, Liliang Ren, Yelong Shen, et al. Phi-4-mini-reasoning: Exploring the limits of small reasoning language models in math. *arXiv preprint arXiv:2504.21233*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Renos Zabounidis, Aditya Golatkar, Michael Kleinman, Alessandro Achille, Wei Xia, and Stefano Soatto. Re-FORC: Adaptive reward prediction for efficient chain-of-thought reasoning. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026.

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. In *Conference on Language Modeling (COLM)*, 2025a.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, pp. 335–340, 2018. doi: 10.1145/3278721.3278779.

Dan Zhang, Min Cai, Jonathan Light, Ziniu Hu, Yisong Yue, and Jie Tang. TDRM: Smooth reward models with temporal difference for LLM RL and inference. *arXiv preprint arXiv:2509.15110*, 2025b.

Jenny Zhang, Shengran Hu, Cong Lu, Robert Lange, and Jeff Clune. Darwin Gödel machine: Open-ended evolution of self-improving agents. *arXiv preprint arXiv:2505.22954*, 2025c. URL https://arxiv.org/abs/2505.22954.

Shu Zhou, Rui Ling, Junan Chen, Xin Wang, Tao Fan, and Hao Wang. When more thinking hurts: Overthinking in LLM test-time compute scaling. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 23967–23977, 2026. doi: 10.18653/v1/2026.findings-acl.1199. URL https://aclanthology.org/2026.findings-acl.1199/.

A Notation Glossary

h_t, $\mathcal{F}_t$, z_t	complete reasoning state; its filtration; observed tap $z_t = f(h_t)$
π_t, $\hat{\pi}_t$	commitment belief (answer-if-stopped-now correct); its probe
Φ_t, $\Phi(h)$, $\hat{\Phi}$	verdict $\mathbb{E}[Y \mid \mathcal{F}_t]$; its state form, equal to Φ_t under Assumption 1 (App. E); its probe
$\Delta_t = b_t + \eta_t$	belief increment = predictable drift + innovation
$m_t = \mathbb{E}[\Delta_t \mid z_t, t]$	tap-conditional drift (what a critic can see)
B_t, $\hat{B}_t$	accumulated drift (Doob compensator); its estimate
$g_\phi(h_t, t)$	drift critic (sensor); $\tilde{\pi}_t = \hat{\pi}_t - \hat{B}_t$ compensated belief
Signal A / B	forced-answer correctness (probe-independent) / probe readout
overthink-wrong, stay-correct	correct at a checkpoint ≤ 2048 then wrong / then right at the end

B Full Audit Matrix

Table 3: Full pre-registered gate and post-hoc robustness matrix; quantities 1.5B / 7B. The pre-registered gates were fixed before the corresponding real-data runs; post-hoc analyses are distinguished below and in the text.

Pre-registered	**outcome**
Gate 0: apparatus calibration	passed (slope 1.04, R^2 0.82; matched floor 0.05)
Gate 1(a): $\geq$ 30 overthink-wrong chains	failed $(10/367) \rightarrow$ committed 7B escalation: $9/262$
Gate 2: recovery at some κ	not met: no tested alarm recovers accuracy; this does not exclude local adverse drift
Verdict rule (viii)	not triggered: verdict AUC 0.747/0.735 does not clear 0.75 (App. H)
Robustness	**outcome**
Forcing operator	of 160 truncations, 69% yield no natural answer; on the rest, agreement 85.7%; greedy vs. sampled $+0.7$ pts $[-0.3, +1.8]$
Checkpoint grid	density inflates overthink-wrong: 5 (1024) = 5 (512) $\rightarrow$ 7 (256) of 240
Label definitions	prevalence 1–4% under all final-outcome variants
Critic capacity	linear critic keeps the signal at 1.5B (R^2 0.16 vs. 0.26)
Alarm threshold	no percentile in $\{50..99\}$ beats full budget (95th: -0.9 $[-2.7, +0.0]$ pts)
Tap richness	ordering holds only at the 1.5B primary tap; unresolved overall
Self-consistency	vote curves inherit survivorship ($+12.0$ within / -12.0 comp.)
Third model	Qwen3-1.7B (non-R1-distilled, Qwen lineage): overthink-wrong 3.0% $[1.7\%, 5.3\%]$; survivorship reproduces
Other base families	R1-Distill-Llama 8B: overthink-wrong 2.2% $[1.1\%, 4.2\%]$, within $+16.6$ / comp. -25.9; Phi-4-mini: 2.2% $[1.1\%, 4.2\%]$, $+9.1$ / -16.6
Decoding temperature	at 0.9: overthink-wrong 1.2% (vs. 2.1%); within-cohort still positive
Synthetic sweep	median recovery 98.6% over 19 configs; failure mode: belief saturation

C Extended Related Work

Probes, values, estimation. Hidden states encode answer correctness (Zhang et al., 2025a) and linearly represent belief-state geometry (Shai et al., 2024)—the license for our readouts: commitments can drift, whereas exact verdicts on nested histories cannot; lossy instantaneous taps and fitted probes need not inherit that null (App. I). Process reward models and MC/TD variants estimate step values (Lightman et al., 2024; Wang et al., 2024; Zhang et al., 2025b); ours is orthogonal.

The training objective is adversarial conditional-moment estimation (Dikkala et al., 2020; Lewis & Syrgkanis, 2018) in the predictability-minimization family (Schmidhuber, 1992; Ganin et al., 2016; Zhang et al., 2018); post-hoc self-revision alone is unreliable (Huang et al., 2024; Madaan et al., 2023).

D Proofs

Theorem 6 (Ballistic bias, diffusive variance)**.** *If $|b_t| \le \beta$ and $|\eta_t| \le c$ a.s., then $\pi_T - \pi_0 = B_T + (M_T - M_0)$ with $B_T = \sum_{t<T} b_t$, $|B_T| \le \beta T$, and (M_t) a martingale with increments bounded by c, so $\Pr(|M_T - M_0| \ge \varepsilon) \le 2\exp(-\varepsilon^2/2Tc^2)$ (Azuma–Hoeffding). If, for a fixed sign $s \in \{-1, 1\}$, $sb_t \ge \beta_0 > 0$ a.s. for every $t < T$, then with probability at least $1 - \alpha$, for $0 < \alpha < 1$,*

$$s(\pi_T - \pi_0) \ge \beta_0 T - c\sqrt{2T \log(1/\alpha)}.$$

This is a finite-horizon statement. On the probability scale the persistence assumption entails $\beta_0 T \le 1$ in expectation; it is not an asymptotic growth law for bounded beliefs. Log-odds drift requires its own decomposition.

Theorem 6. Summing (1), $M_t := \pi_t - \sum_{s<t} b_s$ has $\mathbb{E}[M_{t+1} - M_t \mid \mathcal{F}_t] = \mathbb{E}[\eta_t \mid \mathcal{F}_t] = 0$ and increments bounded by c; Azuma–Hoeffding gives the tail bound. $|B_T| \le \beta T$ is the triangle inequality. The one-sided inequality applied to $s(M_T - M_0)$ gives the displayed signed bound. Taking expectations under $sb_t \ge \beta_0$ gives $\beta_0 T \le \mathbb{E}[s(\pi_T - \pi_0)] \le 1$. □

Ensembling: innovation and shared drift. Averaging k independent belief processes reduces the standard deviation of their averaged innovation by a factor of $\sqrt{k}$ under identical variances, but leaves common expected drift intact. This concerns belief averaging, not nonlinear voting over terminal answers; vote-curve survivorship is audited in App. H.1.

Theorem 7 (Doob)**.** *Any integrable adapted (π_t) decomposes uniquely as $\pi_t = M_t + B_t$ with (M_t) a martingale and (B_t) predictable, $B_0 = 0$; necessarily $B_t = \sum_{s<t} \mathbb{E}[\Delta_s \mid \mathcal{F}_s]$.*

Uniqueness (App. D) makes the construction canonical: with "systematic drift" meaning the predictable part, there is exactly one such object.

Definition 8 (Commitment conservation)**.** *The commitment is* conserved *(dynamically calibrated) iff $B_t \equiv 0$, equivalently $\mathbb{E}[\pi_{t+1} \mid \mathcal{F}_t] = \pi_t$. This is a dynamics condition, not standard probabilistic calibration, a rationality axiom, or a correctness guarantee: even a constant wrong readout satisfies it, whereas useful computation can increase commitment.*

At certainty, $b_t = \mathbb{E}[\pi_{t+1} \mid \mathcal{F}_t] - 1 \le 0$ on $\{\pi_t = 1\}$, strictly if a decrease has positive conditional probability. A conserved bounded commitment cannot leave 1. A sampled correct answer need not imply certainty; deterministic elicitation at a fully specified state does (App. D).

Theorem 7. Existence by the displayed construction. Uniqueness: if $\pi = M + B = M' + B'$ with both decompositions as stated, $N := M - M' = B' - B$ is a predictable martingale with $N_0 = 0$, so $N_{t+1} = \mathbb{E}[N_{t+1} \mid \mathcal{F}_t] = N_t = \cdots = 0$. □

Proposition 2. (i) $\mathbb{E}[\Phi_{t+1} \mid \mathcal{F}_t] = \mathbb{E}[\mathbb{E}[Y \mid \mathcal{F}_{t+1}] \mid \mathcal{F}_t] = \mathbb{E}[Y \mid \mathcal{F}_t] = \Phi_t$ by the tower property, since $\mathcal{F}_t \subseteq \mathcal{F}_{t+1}$; Y is bounded, so all conditional expectations exist. No further hypothesis is used—in particular none on the dynamics, the tap, or the stopping time—so a chain that overthinks has a conserved verdict exactly as one that does not. Its Doob decomposition (Thm. 7) is $\Phi_t = \Phi_t + 0$, so $B \equiv 0$ by uniqueness. (ii) For $s \ge t$, $\mathbb{E}[\pi_{s+1} - \pi_s \mid \mathcal{F}_t] = \mathbb{E}[b_s \mid \mathcal{F}_t]$; summing from $s = t$

to S and letting $S \to \infty$, the commitment side converges by conditional dominated convergence (the chain is frozen after τ, so $\pi_S \to \pi_\tau$ pointwise; $|\pi| \le 1$ and $\pi_\tau = Y$ give $\mathbb{E}[\lim_S \pi_S \mid \mathcal{F}_t] = \mathbb{E}[Y \mid \mathcal{F}_t] = \Phi_t$) and the drift side is dominated by $(\tau - t)_+$ (as $b_s = 0$ for $s \ge \tau$ and $|b_s| \le 1$), integrable because $\mathbb{E}[\tau] < \infty$. This gives $\Phi_t - \pi_t = \mathbb{E}[\sum_{s \ge t} b_s \mid \mathcal{F}_t]$, i.e. (2). If $\pi_t = \Phi_t$ a.s. for all t then (π_t) is a martingale, so $B \equiv 0$ by uniqueness; conversely $B \equiv 0$ makes the right side vanish. Writing $\Phi_t = \Phi(h_t)$ for a measurable Φ requires additionally that h_t be sufficient for $\mathcal{F}_t$ with respect to Y (Assumption 1), which is used nowhere above and is an assumption about the *observation*, not the process. □

Theorem 9 (Convergence and budget-invariance)**.** *A calibrated belief ($B_t \equiv 0$) converges a.s. and in L^1 and, for every stopping time $\tau \le T$, satisfies $\mathbb{E}[\pi_\tau \mid \mathcal{F}_0] = \pi_0$. Thus its expected conclusion is budget-invariant; a biased chain instead adds $\mathbb{E}[B_\tau \mid \mathcal{F}_0]$.*

This concerns the mean, not the full stopped distribution; interior beliefs can decrease on individual paths. Calibration also yields a polarization bound (Thm. 10); adverse drift produces an interior optimal budget in the synthetic model. Proofs are in App. D.

Theorem 9. Convergence: Doob's martingale convergence theorem for bounded martingales; L^1 by uniform integrability. Budget-invariance: optional stopping for bounded martingales ($\tau \le T$). For the biased case apply optional stopping to $M_t = \pi_t - B_t$. □

Certainty at the boundary. If $0 \le \pi_t \le 1$, then on $\{\pi_t = 1\}$ we have $b_t = -\mathbb{E}[1 - \pi_{t+1} \mid \mathcal{F}_t] \le 0$. The inequality is strict wherever $\Pr(\pi_{t+1} < 1 \mid \mathcal{F}_t) > 0$. If $b_t = 0$, the nonnegative variable $1 - \pi_{t+1}$ has conditional expectation zero and therefore vanishes a.s. Iteration makes 1 absorbing for a bounded martingale. The same argument applies to 0. Thus a single stochastic correct elicitation is not proof of certainty, but an exact binary correctness readout cannot undergo a positive-probability correct-to-wrong transition with zero adverse drift at every preceding correct state. With lossy taps or learned probabilities, this implication cannot be inferred from one observed label alone. □

Theorem 10 (Polarization / step count)**.** *Assume calibration, $0 < \sigma^2 \le 1$, and relative innovation non-degeneracy, $\mathbb{E}[\eta_t^2 \mid \mathcal{F}_t] \ge \sigma^2\, \pi_t(1-\pi_t)$ for all t (the right-hand side vanishes at $\pi_t \in \{0, 1\}$, so absorbing endpoints remain admissible). Then $\mathbb{E}[\pi_T(1-\pi_T)] \le (1-\sigma^2)^T/4$. For $0 < \varepsilon \le 1/2$ and $0 < \delta < 1$, after $T = O\big(\sigma^{-2} \log(1/\varepsilon\delta)\big)$ steps the belief has exited to within ε of a definitive verdict with probability $\ge 1 - \delta$, and bounded threshold stopping preserves the mean by Thm. 9. More information per step (larger σ) means fewer steps; the dependence on the target is logarithmic.*

Theorem 10. Let $v_t = \pi_t(1-\pi_t)$. Calibration gives $\mathbb{E}[v_{t+1} \mid \mathcal{F}_t] = v_t - \mathbb{E}[\eta_t^2 \mid \mathcal{F}_t] \le v_t(1-\sigma^2)$ for every t, using $\mathbb{E}[\eta_t^2 \mid \mathcal{F}_t] \ge \sigma^2 v_t$. Iterating, $\mathbb{E}[v_T] \le (1-\sigma^2)^T/4$. On $\{\pi_T \in (\varepsilon, 1-\varepsilon)\}$, $v_T \ge \varepsilon(1-\varepsilon) \ge \varepsilon/2$, so Markov's inequality gives $\Pr(\pi_T \in (\varepsilon, 1-\varepsilon)) \le (1-\sigma^2)^T/(2\varepsilon) \le \delta$ for $T \ge \sigma^{-2} \log(1/(2\varepsilon\delta))$. Validity of threshold stopping is Thm. 9. □

Accuracy is not posterior polarization. For a calibrated posterior p_t on a fixed binary hypothesis, the Bayes decision accuracy is $\mathbb{E}[\max\{p_t, 1-p_t\}]$, which can increase while $\mathbb{E}[p_t]$ is constant. For an answer-if-stopped-now success probability π_t, expected accuracy is instead $\mathbb{E}[\pi_t]$. These are different readouts. In particular, a martingale decomposition alone does not justify an accuracy expansion of the form $a + c\sqrt{T} - dT$, or an interior optimal budget. The interior peak in Sec. 5.1 is a property of the specified synthetic decision model, measured numerically. Equation (4) is the corresponding exact statement for operational commitment accuracy.

Proposition 15. $\mathbb{E}[\Delta_t^2] = \mathbb{E}[(b_t + \eta_t)^2] = \mathbb{E}[b_t^2] + \mathbb{E}[\eta_t^2]$ since the cross term vanishes by the tower property (b_t is $\mathcal{F}_t$-measurable). A penalty $\rho(\Delta_t)$ with ρ locally quadratic charges both terms at the same second-order rate. Consecutive-step KL between readout distributions is, to second order, at interior probabilities, the Fisher-weighted squared increment $\Delta_t^2/(2\pi_t(1-\pi_t))$; the weight is $\mathcal{F}_t$-measurable, so $\mathbb{E}[w\Delta_t^2] = \mathbb{E}[w\, b_t^2] + \mathbb{E}[w\, \eta_t^2]$ with the same weight on both components. This is a local expansion away from the endpoints, not a global comparison between KL and squared increments. □

Lemma 16. $2\,\mathbb{E}[g\Delta] - \mathbb{E}[g^2] = 2\,\mathbb{E}[g\, m] - \mathbb{E}[g^2] = \mathbb{E}[m^2] - \mathbb{E}[(g-m)^2]$ using the tower property ($\mathbb{E}[g(h_t)\Delta_t] = \mathbb{E}[g(h_t)m_t]$), maximized at $g = m$ over any class containing m; for general $\mathcal{G}$ the display equals $\mathbb{E}[m^2] - \mathbb{E}[(g-m)^2]$, whose supremum over $\mathcal{G}$ is $\mathbb{E}[m^2] - \inf_{g \in \mathcal{G}} \mathbb{E}[(g-m)^2]$. □

Theorem 18. (i) Sum Lemma 16 over t (the critic receives the step index, so the sum is attained term-wise); under Assumption 1, $m_t = b_t$. (ii) If $m_t = 0$ then $\mathbb{E}[g\Delta_t] = \mathbb{E}[g\, m_t] = 0$ for every $g \in \mathcal{G}$, so the inner objective is $-\mathbb{E}[g^2] \le 0$, maximized at $g \equiv 0$ with value 0. □

Proposition 11 (Graceful degradation)**.** *Under Assumption 1, if $|g_\phi(h_t, t) - m_t| \le \delta$, then for any stopping time τ with $\mathbb{E}[\tau] < \infty$, $\left|\mathbb{E}[\tilde{\pi}_\tau - \tilde{\pi}_0]\right| \le \delta\, \mathbb{E}[\tau]$, and conditionally $\left|\mathbb{E}[\tilde{\pi}_\tau \mid \mathcal{F}_0] - \tilde{\pi}_0\right| \le \delta\, \mathbb{E}[\tau \mid \mathcal{F}_0]$.*

Proof. $\tilde{\pi}_{t+1} - \tilde{\pi}_t = \Delta_t - g_\phi(h_t, t)$ has conditional mean $b_t - g_\phi(h_t, t) = m_t - g_\phi(h_t, t) \in [-\delta, \delta]$ under Assumption 1. Write $\tilde{\pi}_t = N_t + D_t$ with (N_t) a martingale ($N_0 = \tilde{\pi}_0$, increments equal to $\Delta_t - b_t$ and bounded by 2) and (D_t) predictable with $D_0 = 0$ and $|D_{t+1} - D_t| \le \delta$. Optional stopping for bounded-increment martingales at τ with $\mathbb{E}[\tau] < \infty$ gives $\mathbb{E}[N_\tau] = \mathbb{E}[N_0] = \mathbb{E}[\tilde{\pi}_0]$, and $|D_\tau| \le \delta\tau$ gives $|\,\mathbb{E}[\tilde{\pi}_\tau - \tilde{\pi}_0]| = |\,\mathbb{E}[D_\tau]| \le \delta\, \mathbb{E}[\tau]$; conditioning on $\mathcal{F}_0$ throughout gives the second form. □

E The Verdict in Operator Form

Let $(Pf)(h) = \mathbb{E}[f(h_{t+1}) \mid h_t = h]$ (Assumption 1) and $L = P - I$.

Proposition 12 (Operator form)**.** *(a) $b_t = (L\pi)(h_t)$: drift is the generator applied to the readout. (b) The chain is calibrated iff $P\pi = \pi$, i.e. the readout is* harmonic*: $\pi \in \ker L$. (c) Assume the terminal structure is* absorbing*: from every interior state the boundary (an emitted answer, carrying $g_\partial = \mathbf{1}\{\textit{correct}\}$) is reached with probability one. Then Φ is the unique* bounded *solution of the discrete Dirichlet problem $L\Phi = 0$ with boundary data g_∂, and $\Phi(h) = \mathbb{E}_h[g_\partial(h_\tau)] = V(h)$: the verdict is the harmonic extension of terminal correctness, i.e. the exact value function. Under Assumption 1 this state function agrees with the filtration-conditional verdict of Prop. 2, $\Phi(h_t) = \mathbb{E}[Y \mid \mathcal{F}_t] = \Phi_t$ a.s., which is what licenses writing Φ as a function of the state at all. (d) $\sum_t \mathbb{E}[b_t^2] = \|(P - I)\pi\|^2_{L^2(\mu)}$, where $\mu(A) = \mathbb{E}\left[\sum_{t<\tau} \mathbf{1}\{h_t \in A\}\right]$ is the* unnormalized *pre-absorption occupancy measure (no horizon factor is needed with this normalization): the training objective of Sec. 4 is a squared generator residual, minimized exactly on harmonic readouts on the occupied states. It is not the usual Dirichlet form $\langle \pi, (I - P)\pi \rangle$.*

Proof. (a), (b) are definitional; (d) sums pointwise. For (c), define $\Phi(h) = \mathbb{E}_h[g_\partial(h_\tau)]$. Almost-sure absorption makes this bounded candidate well-defined; the Markov property gives $P\Phi = \Phi$ in the interior and its boundary values are g_∂. Uniqueness holds because if u is bounded, harmonic, and zero on the boundary, then $u(h) = \mathbb{E}_h[u(h_{t\wedge\tau})] \to 0$ by dominated convergence under a.s. absorption. Without absorption uniqueness genuinely fails: $u(h) = \Pr_h(\tau = \infty)$ is bounded, harmonic, and vanishes on the boundary. □

Fundamental-matrix form (finite case). If the interior state space is finite (e.g. a finite discretization, or the state augmented with a step index under a hard budget $T_{\max}$, which makes the interior block nilpotent) and the chain is absorbing, write $P = \left(\begin{smallmatrix} Q & R \\ 0 & I \end{smallmatrix}\right)$. Then $Q^t \to 0$, hence $\rho(Q) < 1$, $I - Q$ is invertible, and

$$\Phi_{\text{int}} = (I - Q)^{-1} R\, g_\partial, \qquad (I - Q)^{-1} = \sum_{t \ge 0} Q^t \tag{5}$$

(Kemeny & Snell, 1976, Thm. 3.2.1): $N = (I - Q)^{-1}$ is the fundamental matrix (the Green's operator of L), $N_{hh'}$ the expected visits to h' before absorption, and NR the absorption-distribution map. On infinite state spaces sub-stochasticity does *not* imply $\rho(Q) < 1$; the probabilistic representation of Prop. 12(c) is the correct general statement, and (5) has a probabilistic interpretation through visit counts. Almost-sure absorption alone does not ensure finite expected duration or a bounded inverse on an infinite state space.

Poisson representation and horizon sensitivity. Writing $d = \pi - \Phi$ on the transient states, the common terminal boundary gives $(I - Q)d = -b$, hence $d = -Nb$. Consequently

$$|d(h)| \le \mathop{\mathbb{E}}_h\left[\sum_{t<\tau} |b(h_t)|\right] \le \|b\|_\infty \mathop{\mathbb{E}}_h[\tau].$$

Expected remaining duration therefore controls how local drift translates into a commitment–verdict gap. In the finite absorbing case $Q^t \to 0$; if Q is diagonalizable, convergence is bounded by a constant times $\rho(Q)^t$. Diagonalizability of an arbitrary Markov operator alone is insufficient: other unit-modulus eigenvalues can prevent convergence.

On log-odds "closed forms." The compensated belief $\tilde{\pi}_t = \pi_t - \hat{B}_t$ estimates the *martingale component* M_t (Thm. 7); it differs from the verdict Φ_t by the expected *total* accumulated drift $\mathbb{E}[B_\infty \mid \mathcal{F}_t]$, while the raw belief differs from Φ_t by expected *future* drift alone (2). No identity of the form $\Phi = \sigma(\langle u_A, h\rangle - \sum_{t\geq s} b_t)$ holds: expectations do not commute with σ, and probability-scale drift is not a log-odds increment. If a log-odds compensation is wanted, define the log-odds drift $\tilde{b}_t = \mathbb{E}[L_{t+1} - L_t \mid \mathcal{F}_t]$ for $L_t = \operatorname{logit} \pi_t$ and compensate inside the logit; we use the probability scale throughout.

F The Variance-Blind Objective: Formal Statements

This appendix states and discusses the results summarized in Sec. 4. In the critic formulas, h_t denotes the available input (the tap z_t in Sec. 3.1), with time included when supplied. Equality $m_t = b_t$ requires a sufficient input; otherwise the objectives target only the predictable component visible to that input.

Proposition 13 (One-sided erosion objective)**.** *For a square-integrable increment Δ_t, let $m_t = \mathbb{E}[\Delta_t \mid z_t, t]$ and $m_t^- = \max\{-m_t, 0\}$. Then*

$$\sup_{g(z_t,t)\leq 0} \{2\,\mathbb{E}[g\Delta_t] - \mathbb{E}[g^2]\} = \mathbb{E}[(m_t^-)^2], \qquad g^\star = \min\{m_t, 0\},$$

where the supremum is over square-integrable, measurable functions. The value is zero iff $m_t \geq 0$ a.s.; thus it penalizes adverse predictable movement without directly penalizing beneficial predictable movement or residual variance.

Proof. Conditioning gives $2\,\mathbb{E}[g\Delta_t] - \mathbb{E}[g^2] = \mathbb{E}[2gm_t - g^2]$. Pointwise, the concave quadratic is maximized over $g \leq 0$ at $\min\{m_t, 0\}$, with value $(m_t^-)^2$. This maximizer is square-integrable by conditional Jensen. A restricted critic class containing zero has a nonnegative supremum no larger than this value, with equality when it contains the maximizer. □

This is a derived alternative to the symmetric objective, not a reported training experiment. Under a sufficient tap, its zero set is the submartingale condition $b_t \geq 0$, which by Cor. 4 excludes an accuracy-improving stopping rule under the same protocol. Estimator error, representation shift, and task loss still matter during optimization.

Proposition 14 (Readout error propagates into measured drift)**.** *Let $\hat{\pi}_t = \pi_t + e_t$ be adapted with square-integrable errors, and define $\hat{b}_t = \mathbb{E}[\hat{\pi}_{t+1} - \hat{\pi}_t \mid \mathcal{F}_t]$. Then*

$$\hat{b}_t - b_t = \mathbb{E}[e_{t+1} \mid \mathcal{F}_t] - e_t, \qquad \|\hat{b}_t - b_t\|_2 \leq \|e_{t+1}\|_2 + \|e_t\|_2.$$

If $|e_t| \leq \epsilon$ a.s. at every step, the pointwise drift error is at most 2ϵ. The same bound holds after projection onto a tap.

Proof. Subtract the conditional increment definitions and apply conditional expectation contraction in L^2, followed by the triangle inequality. The uniform bound follows directly; another conditional expectation is also a contraction. □

Neither AUC nor binned ECE supplies such a uniform error bound. Consequently, predictable increments of a learned probe certify predictability of that probe, not automatically of the underlying behavioral success probability.

The objective, as used in the main text. Penalizing the variational value of Sec. 4 gives the label-free, variance-blind objective

$$\min_{\theta} \ \mathcal{L}_{\text{task}}(\theta) + \lambda \max_{\phi} \sum_{t} \big(2\,\mathbb{E}[g_\phi(h_t,t)\Delta_t] - \mathbb{E}[g_\phi(h_t,t)^2]\big).$$

With a fixed readout and a critic class containing m_t, the inner maximum is exactly $\sum_t \mathbb{E}[m_t^2]$: it vanishes iff the chain is state-calibrated, while admissible zero-mean innovations incur no direct penalty.

Proposition 15 (Locally quadratic penalties charge both components)**.** *Since b_t is $\mathcal{F}_t$-measurable, $\mathbb{E}[\Delta_t^2] = \mathbb{E}[b_t^2] + \mathbb{E}[\eta_t^2]$. Any penalty that is locally quadratic in the per-step belief change—squared-increment regularizers and, to second order, KL-type proximity penalties between consecutive* readout *distributions—charges innovation on the same footing as drift. It therefore penalizes useful unpredictable updates as well as predictable movement; the effect of optimization on either component depends on the model and task loss.*

We penalize only the predictable part. As a conditional expectation, it admits a variational characterization—the classical L^2 projection, used here as an *adversarial conditional-moment* estimator (Dikkala et al., 2020; Lewis & Syrgkanis, 2018).

Lemma 16 (Variational identity)**.** *Let $m_t = \mathbb{E}[\Delta_t \mid h_t]$ and let $\mathcal{G}$ be a function class containing m_t. Then*

$$\mathbb{E}[m_t^2] = \max_{g \in \mathcal{G}} \big\{2\ \mathbb{E}[g(h_t)\,\Delta_t] - \mathbb{E}[g(h_t)^2]\big\}, \tag{6}$$

attained at $g^\star = m_t$. For a capacity-limited $\mathcal{G}$, the supremum equals $\mathbb{E}[m_t^2] - \inf_{g\in\mathcal{G}} \mathbb{E}[(g - m_t)^2]$, the representable share of the drift energy: guarantees hold against drift representable by the critic class.

Definition 17 (Drift-penalized training)**.** *With a frozen readout and a critic $g_\phi(h_t, t)$ (state and step index),*

$$\min_{\theta} \ \mathcal{L}_{\text{task}}(\theta) \ + \ \lambda \max_{\phi} \sum_{t} \big(2\ \mathbb{E}[g_\phi(h_t,t)\,\Delta_t] - \mathbb{E}[g_\phi(h_t,t)^2]\big). \tag{7}$$

The critic forecasts the next belief move; the reasoner is trained to make its own updates unforecastable. The penalty uses no gold labels.

Theorem 18 (Identification of drift energy and variance-blindness)**.** *(i) Provided $\mathcal{G}$ contains the map $(h,t) \mapsto m_t(h)$, the inner maximum of (7) equals $\sum_t \mathbb{E}[m_t^2]$; it is zero iff the chain is state-calibrated ($m_t \equiv 0$), and under Assumption 1 equals $\sum_t \mathbb{E}[b_t^2]$ with zero iff $B_t \equiv 0$. (ii)* Variance-blindness*: if $m_t = 0$ with η_t arbitrary, the optimal critic is $g \equiv 0$ and the penalty vanishes. The objective does not directly penalize zero-mean innovations; this orthogonality does not guarantee that optimizing a shared model preserves accuracy, information acquisition, or innovation variance.*

Algorithm 1 (App. G) instantiates (7) with a two-timescale update. Freezing the head fixes its weights but does not fix its input representation or calibration. Equation (6) descends from predictability minimization and adversarial conditional-moment estimation (lineage in Sec. 6); what is new is the semantics—the value held fixed, the *dynamics* regularized, the variance-blindness guarantee (ii), and the critic's second job at inference (Sec. 4). Sec. 5.1 validates the mechanism where ground truth is available; fine-tuning a full reasoning model with (7) is future work (App. G).

G Deployment Notes and Future Training Work

Serving (Level 1). One critic forward pass per step; maintain $\hat{B}_t = \sum_{s<t} g_\phi(h_s, s)$; alarm when $-\hat{B}_t > \kappa$; answer from the last safe checkpoint (or force an answer at the trigger point). The compensated belief $\tilde{\pi} = \hat{\pi} - \hat{B}$ is the drift-purged reading used for answer selection. κ is set from the erosion distribution of healthy (stay-correct) held-out chains.

Level 2 (state steering). Let $J_t = (\nabla_h g_\phi(h_t, t))^\top$ be the $1 \times d$ Jacobian row of the scalar critic. Then $u_t = -(J_t^\top J_t + \rho I)^{-1} J_t^\top g_\phi(h_t, t)$, with $J_t^\top J_t$ and I both $d \times d$, is the Gauss–Newton step for $\min_u \|g_\phi(h_t + u, t)\|^2$ to first order; it vanishes as $g_\phi \to 0$. For a *linear* critic J_t does not depend on

the current hidden state, so the update follows a time-indexed fixed direction with state-dependent magnitude—the sense in which static activation steering (Li et al., 2023) is the degenerate case. A critic constant in h_t has $J_t = 0$ and produces no steering at all. Variance-gate the intervention with the critic's uncertainty head; fall back to Level 1, which does not modify hidden states.

Implementation choices for language-model training (Alg. 1). The θ-step differentiates the penalty through generated text. Sampled tokens are discrete, so the pathwise gradient is unavailable as written and a concrete estimator must be chosen: policy-gradient/REINFORCE on the sampled trajectory, a teacher-forced fixed-token surrogate, a straight-through relaxation, or differentiable latent reasoning. Our synthetic study differentiates a continuous belief directly and so does not settle this; the reference implementation we release adopts the teacher-forced surrogate (sample a rollout without gradient, re-forward it with gradient, and propagate only through the readout's dependence on the hidden states). We therefore state Alg. 1 as conceptual for autoregressive models.

A frozen head is not a frozen ruler. Freezing the readout's *parameters* prevents direct updates to the readout weights, but updating θ changes the representation h_t that feeds the readout, so its calibration and semantics can drift even with fixed weights. Any real fine-tuning run should therefore monitor readout calibration on held-out problems during training, or freeze the tapped representation below the intervention point; we do not consider this solved.

Feedback trap. Never fit the critic on corrected trajectories: the controller removes the signal the critic estimates. Audit rollouts must stay open-loop.

Algorithm 1 Drift-penalized training (two-timescale; conceptual—see App. G for the gradient estimator a language model requires)

train readout r on base-model rollouts; **freeze** r; $\pi_t = r(h_t)$
while not converged **do**
  roll out uncorrected chains; $\Delta_t^{(i)} = \pi_{t+1}^{(i)} - \pi_t^{(i)}$
  for $k = 1$ to K **do**
    $\phi \mathrel{+}= \alpha_\phi \nabla_\phi \frac{1}{n} \sum_{i,t} \left(2 g_\phi(h_t^{(i)}, t)\Delta_t^{(i)} - g_\phi(h_t^{(i)}, t)^2\right)$
  end for
  $\theta \mathrel{-}= \alpha_\theta \nabla_\theta(\widehat{\mathcal{L}_{\text{task}}} + \lambda \widehat{V})$ (ϕ, r fixed)
end while
return θ; retain g_ϕ as the inference-time drift sensor

Algorithm 2 Drift audit / critic fit (frozen model θ)

roll out uncorrected chains on $\{q^{(i)}\} \sim \mathcal{D}$; record $h_t^{(i)}$, $\pi_t^{(i)} = r(h_t^{(i)})$
fit $g_\phi = \arg\min_g \sum_{i,t} \left(\Delta_t^{(i)} - g(h_t^{(i)}, t)\right)^2$ (equivalent to (6))
return g_ϕ; report held-out $\widehat{V} = \widehat{\mathbb{E}}[2g_\phi\Delta - g_\phi^2]$ and chain-clustered uncertainty

Training with gold anchors (future work). Eq. (7) enforces martingale (self-consistency) calibration; it is silent about *which* verdict is conserved. When verified $(q, a^\star)$ pairs exist, a complementary phase teacher-forces the gold answer and supervises the belief toward the harmonic extension of the *verified* boundary with a proper scoring rule—supervising calibration, not token-likelihood. Critic data must remain free-running (the guard above); gold labels are load-bearing only at the boundary. Fine-tuning a full reasoning model with the combined objective, and the accuracy-vs-budget flattening it predicts, is the natural next experiment; we did not run it. Expected pitfalls, for those who do: (i) the frozen-ruler guard matters most under fine-tuning, since the readout would otherwise drift with the model; (ii) interaction with preference-tuned policies is untested—a KL-to-reference term and the drift penalty regularize different objects and should compose, but the two-timescale inner maximization adds instability risk; (iii) rationale style may shift (the penalty discourages predictable belief movement, not predictable *text*); and (iv) with positive average predicted drift at these scales, the penalty's measurable effect may be small until adverse-drift regimes (longer budgets, weaker models, distribution shift) are targeted deliberately.

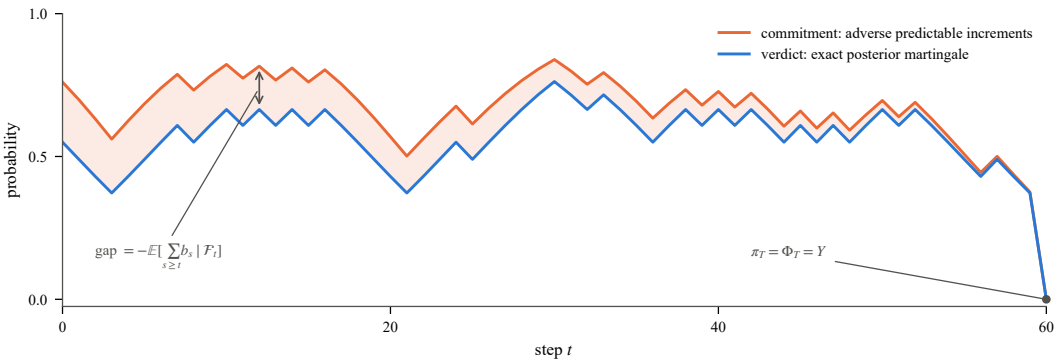


Figure 3: The commitment–verdict gap on a simulated binary-evidence chain with adverse predictable drift (illustrative). The verdict is the exact posterior of a binary terminal outcome, revealed at termination. The commitment adds a nonnegative offset whose conditional expectation decreases, giving adverse predictable drift. The shaded gap equals $-\mathbb{E}[\sum_{s\geq t} b_s \mid \mathcal{F}_t]$ and vanishes when both readouts equal Y. It need not shrink monotonically on an individual path.

H Experimental Details

Internal pre-registration. Gates, thresholds, operating points, label definitions, and interpretation rules were written down and dated *before* the corresponding real-data results were computed, and are reproduced in App. H; the dated plan itself ships with the supplement. The original gates were fixed before any real-data run, and the secondary analyses before any Exp. 1/2 number was unblinded. The record is an internal, dated plan rather than an entry in an external registry; we use "pre-registered" in that sense throughout. Its content is auditable against the reported outcomes: Gate 1(a) failed and triggered its committed fallback, Gate 2 was met at no operating point, and rule (ix) suppressed a detection claim—all reported as such.

Historical cross-protocol branching diagnostic. The earlier study uses the same 12 wrong-ending 1.5B chains at their first correct checkpoint ($\leq$ 2048; all five levels), but samples 16 immediate answers and 16 continuations with *4096 additional tokens and no terminal forcing*. Its continuation success probability Ψ_t is not the primary-policy verdict Φ_t. It gives $\hat{\pi}_t = 0.80$, $\hat{\Psi}_t = 0.16$ and gap $+0.64$; 76% of 192 continuations are capped without an answer, versus 7% of 480 in the earlier 30-prefix control arm. Low success therefore largely reflects nonresolution. The post-hoc partition ($\hat{\pi}_t \geq 0.75$, gap ≥ 0.5) is descriptive. These historical results are retained below but are not substituted for the matched follow-up in Sec. 5.3.

problem	checkpoint	$\hat{\pi}_t$	$\hat{\Psi}_t$	gap
326	2048	1.00	0.00	+1.00
101	1536	1.00	0.00	+1.00
493	1536	1.00	0.19	+0.81
128	512	0.81	0.00	+0.81
274	2048	0.94	0.12	+0.81
120	1536	1.00	0.25	+0.75
126	1024	0.62	0.00	+0.62
398	512	0.94	0.31	+0.62
381	512	0.56	0.00	+0.56
9	2048	0.38	0.00	+0.38
382	2048	0.38	0.06	+0.31
238	1536	0.94	1.00	-0.06

Experiment 0: late-onset drift and the configuration sweep. A pre-registered late-onset test (Fig. 9) makes the penalty load-bearing: a lean beginning only after the supervised horizon survives task supervision, is removed by the drift penalty (85.4% accuracy with gain intact), and defeats the symmetric penalty, which instead saturates the belief. Across 19 synthetic configurations, median

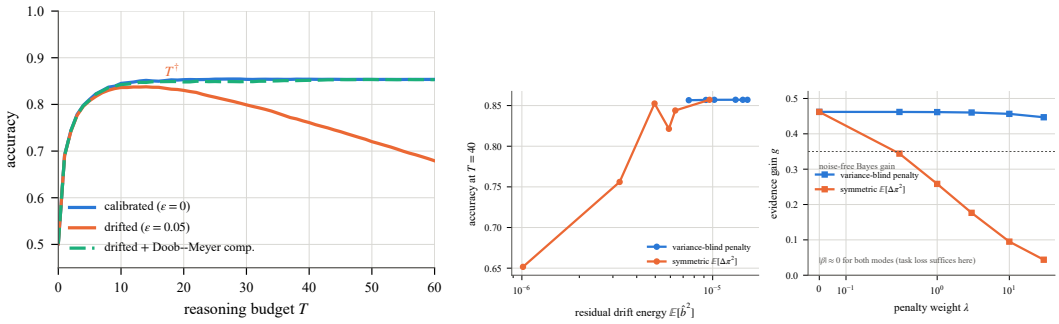


Figure 4: **Left:** synthetic overthinking under decaying evidence and a persistent lean; compensation recovers the calibrated plateau. **Right:** the symmetric penalty destroys evidence gain and accuracy, while the variance-blind penalty leaves both intact across λ.

compensation recovery is 98.6%; the main failure mode is belief saturation, where probability-scale drift becomes hard to observe.

Hardware and stack. One RTX 4070 Ti (12 GB), Windows 11, PyTorch cu126, Hugging Face `transformers` (SDPA attention), fp16. Primary-audit seed 0; follow-up studies use their recorded, disjoint seed streams.

Rollouts. MATH-500 test split; prompt = problem + "*Please reason step by step, and put your final answer within* `\boxed{}`." via the chat template (the R1-distill template opens `<think>` automatically); temperature 0.6, top-p 0.95, max 4096 new tokens, one chain per problem, batch 8.

Forcing. At checkpoints $\{512k\}_{k=1}^{8}$ (truncating mid-think): append `\n</think>\n\n**Final Answer:** \boxed{` and decode greedily ($\leq$ 48 tokens, parsed with the permissive brace-closing rule, which accepts a continuation that never closes); if the chain had already closed its think block, the forcing string is appended without the closing tag. The natural termination is its own checkpoint (the chain's own boxed answer where present). Verification: normalized string match with `math-verify` fallback; unparseable $\Rightarrow$ incorrect.

States and steps. Teacher-forced prefill; hidden states at layers 14/21/27 (of 28) at paragraph boundaries (`\n\n` splits, segments < 10 tokens merged forward, ≤ 96 steps/chain) and at every checkpoint position. Layer 21 is primary; 14/27 form the robustness panel.

Probes and critic. Commitment probe: logistic regression on checkpoint states with label = forced-answer correctness at that checkpoint; C selected by inner CV over $\{0.01, 0.1, 1\}$; temperature-scaled on one fifth of the probe split; 15-bin ECE reported. Verdict probe: identical protocol, label = final-chain correctness. Critic: LayerNorm $\to$ 2$\times$256 GELU MLP, dropout 0.1, mean and log-variance heads, Gaussian NLL, AdamW (10^{-3}, wd 10^{-4}), batch 1024, early stopping; input $(h_t, t/4096)$; targets $\Delta\hat{\pi}_t$ on uncorrected rollouts of the critic split only. Splits: 40/30/30 problem-level, stratified by (level, final correctness).

Labels. *overthink-wrong*: correct at some checkpoint ≤ 2048 tokens (strictly before the final answer) and wrong at the final answer; *stay-correct*: the same but correct finally; chains ending before the first checkpoint have no early checkpoint and can receive neither label; *late-bloomer*: no early correct checkpoint, correct finally; *early-unresolved/wrong*: no early correct checkpoint, wrong finally. The latter is stored as `never_correct` in the archived code and shown as "never-correct" in legacy figures. It does not imply correctness nowhere: the implemented label does not inspect transient correctness after token 2048.

Conventions. The theory's abstract "step" is instantiated as a paragraph segment of the think text (typically 40–90 per chain); checkpoints are token-indexed, and the critic receives the token position, so the two clocks are linked through (h_t, t). "Mid-chain" is the first paragraph step at or after 50%

of the chain's generated length (capped at the last step). Chains with fewer than four paragraph steps are excluded from step-level analyses.

Probe health at the pre-registered tap. Rule (viii) is not triggered at either scale, on its pre-registered AUC condition alone: the verdict probe's final-correctness AUC is 0.747 at 1.5B and 0.735 at 7B, neither above 0.75. Separately—and as a post-hoc judgment, not part of the dated rules—the 1.5B verdict probe is poorly calibrated (ECE 0.138 against 0.074 for the commitment probe), so we do not read its predictable fraction as evidence about the tap.

Probe-shift and difficulty controls. Drift energy near vs. far from probe-training checkpoints is comparable (1.5×10^{-3}/1.6×10^{-3}; 7B 7.0×10^{-3}/7.1×10^{-3}), which does not reveal a large checkpoint-proximity effect; on levels 1–2 (1.5B) both overthink rate and matched-position drift energy fall (1.5% vs. 2.7%; 1.5×10^{-3} vs. 1.7×10^{-3}).

Mean predicted drift by group. 1.5B: $+0.0144$ overthink-wrong, $+0.0052$ stay-correct, $+0.0035$ late-bloomer, $+0.0087$ early-unresolved/wrong; 7B: $+0.0228$, $+0.0154$, $+0.0203$, $+0.0240$.

Per-group outcomes at the fixed operating point (Fig. 5). Alarm at the pre-registered κ (90th stay-correct percentile, critic split), eval levels 3–5: overthink-wrong 0.0% (0.0% strict), stay-correct 97.5%, late-bloomer 83.3%, early-unresolved/wrong 0.0%; fire rates 0% (overthink-wrong) vs. 13% (stay-correct); mean charged tokens 2624 vs. 2808 at full budget.

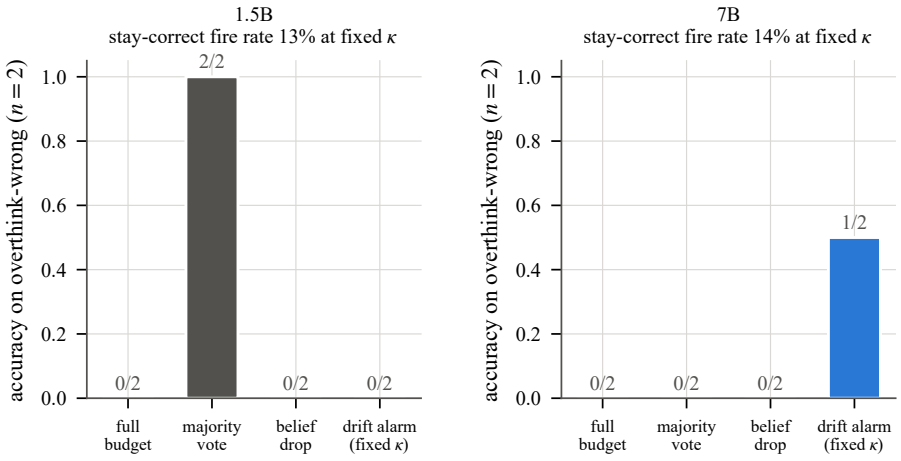


Figure 5: Per-overthink-wrong-chain outcomes at the pre-registered operating point (1.5B left, 7B right; k/n annotated, $n = 2$ eval events per scale). Anecdotal at this prevalence—deferred here per the interpretation rules; fire rates 0% / 50% on overthink-wrong.

Exp 2 policies. All replays are nonanticipating; answers only at recorded checkpoints (the last checkpoint at or before the trigger). $\kappa \in \{50, 70, 80, 90, 95\}$th percentiles of $\max_t(-\hat{B}_t)$ on stay-correct critic-split chains; belief-drop thresholds $\delta \in \{0.1, \ldots, 0.5\}$ for both raw $\hat{\pi}$ and compensated $\hat{\pi} - \hat{B}$; fixed stops at every checkpoint. Tokens are charged by rule (vi) below.

Escalation outcome. Gate 1(a) failed at 1.5B (10 overthink-wrong chains on levels 3–5, minimum 30); the escalation to 7B-NF4 on levels 4–5—the pre-registered fallback of Gate 1(a)—was executed with identical protocol and returned 9 overthink-wrong of 262. Gate 0 passed (Sec. 5.1); Gate 2 was met at no κ at either scale (the alarm loses accuracy at both). Both audits are reported in full. The rule-(ix) predictor comparison is unpowered—the evaluation split holds 2/2 overthink-wrong chains (1.5B/7B) against 79/50 stay-correct, below the pre-committed minimum of three positives for reporting AUC—so no AUC is reported and the raw paired-bootstrap differences are released with the result files in the supplement. The 1.5B stopping-policy replay (full 77.3% at

2808 tokens; alarm 74.5% at 2624; vote 71.8%; agreement 65.5% at 1625) mirrors the 7B replay of Sec. 5.5 (Pareto fronts: Fig. 7 at 7B, Fig. 8 at 1.5B); the 1.5B commitment-probe and compensator trajectories by group are in Fig. 6.

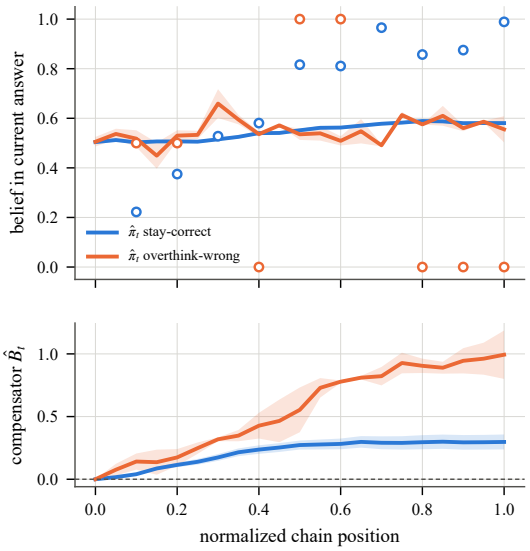


Figure 6: 1.5B belief and compensator trajectories by group (eval, levels 3–5; illustrative—two eval overthink-wrong chains).

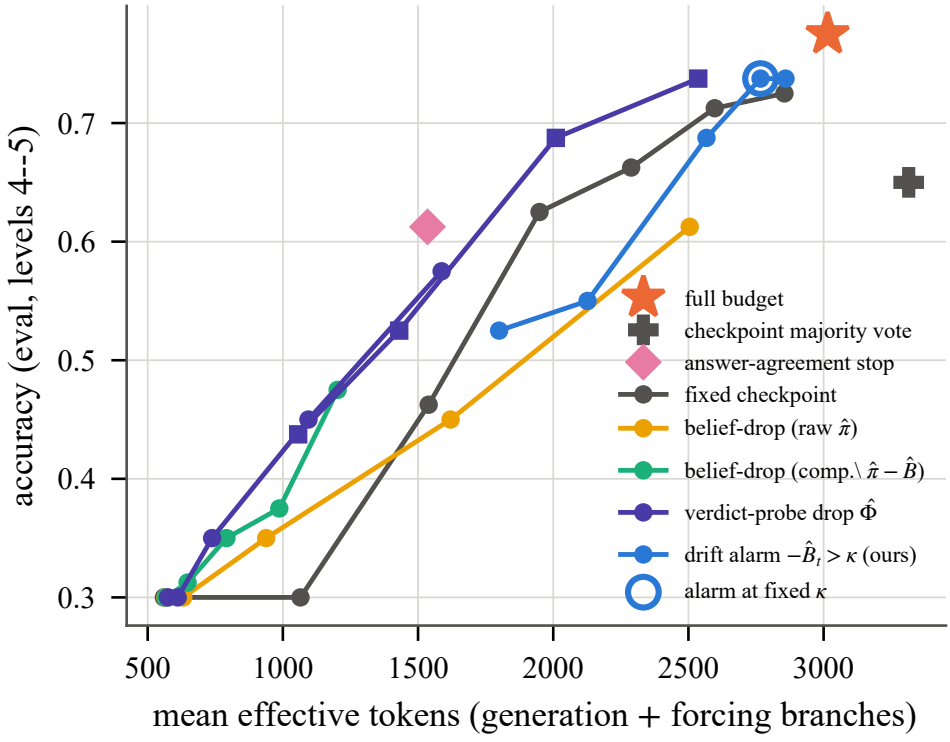


Figure 7: Stopping-policy replays at 7B (levels 4–5 eval): every family sits below the full-budget point; the monotone cohort curve separately explains why *global truncation* has no favorable operating point.

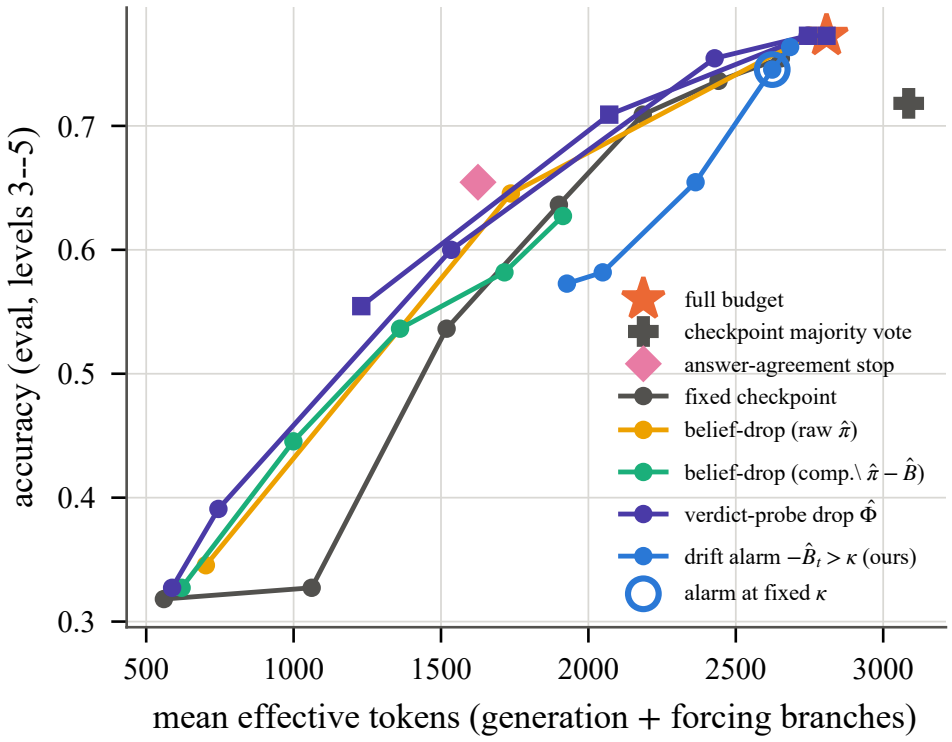


Figure 8: 1.5B stopping-policy replay (levels 3–5 eval), analogous to Fig. 7.

Gates (pre-committed). Gate 0: sensor slope in $[0.9, 1.1]$, $R^2 > 0.8$, no false drift at $\varepsilon = 0$, detection floor reported. Gate 1: (a) ≥ 30 overthink-wrong chains on levels 3–5, with the committed fallback "if not, escalate to 7B-4bit on levels 4–5 with identical protocol before concluding absence"; (b) AUC$(-\hat{B}_{\text{mid}}) \geq 0.65$ and within 0.02 of the PCA-32 baseline or better; (c) onset localization consistent with the flip. Gate 2: at some κ, recovery of $\geq 50\%$ of the overthinking loss on overthink-wrong chains; stay-correct accuracy change ≥ -1 pt with fire rate $\leq 10\%$; tokens saved $\geq 15\%$ at matched-or-better overall accuracy.

Secondary pre-registered analyses (fixed before any Exp. 1/2 number was unblinded). The original gates above are unchanged. The following analyses, definitions, and decision rules were fixed while the rollouts were running and before any real-data result was computed. *(i) Operating point.* The reported alarm operating point is fixed in advance as the 90th percentile of $\max_t(-\hat{B}_t)$ on stay-correct critic-split chains; other percentiles appear only in the Pareto sweep. *(ii) Strict labels.* Because a single forced answer is a noisy draw, all headline quantities are additionally reported under a strict label: overthink-wrong-strict requires correctness at two consecutive checkpoints ≤ 2048 with the same normalized answer string, and a wrong final answer; stay-correct-strict analogously. The population and the conclusions must survive this tightening. Per-checkpoint answer consistency (fraction of checkpoints agreeing with the chain's modal answer) is reported by group. *(iii) Baseline gauntlet (Exp 1).* The predictor comparison additionally includes: a supervised MLP (critic architecture) classifying overthink-wrong vs. stay-correct from h_{mid}; a length-only predictor; a Signal-A-only predictor (forced-answer pattern up to mid-chain); an entrenchment score in the style of He et al. (2025) (per-chain regression of $\Delta\hat{\pi}$ on $\hat{\pi}$); and critic ablations (no step-index input; input $(\hat{\pi}_t, t)$ only). Paired bootstrap confidence intervals are reported for AUC$(-\hat{B})$ − AUC(probe slope) and AUC$(-\hat{B})$ − AUC(supervised MLP). *(iv) Prediction vs. postdiction.* All predictor rows are also reported restricted to overthink-wrong chains whose forced answer is still correct at mid-chain. *(v) Policy gauntlet (Exp 2).* Added causal-replay policies: checkpoint majority vote over the eight forced answers; answer-agreement stopping (two consecutive checkpoints with the same normalized answer); verdict-probe level and drop stopping; the fixed stop at 2048 called out against Gate 2's region; and alarm vs. raw-drop detection compared at matched stay-correct fire rate. *(vi) Token ac-*

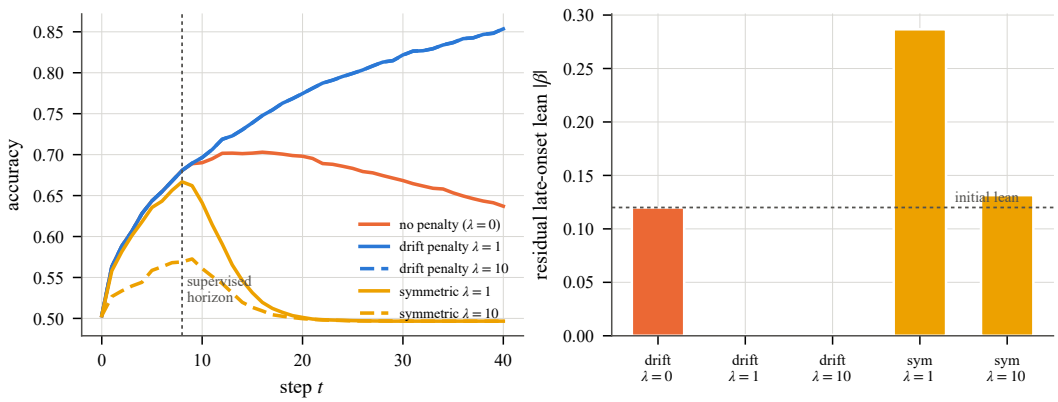


Figure 9: Late-onset lean: the lean acts only after the supervised horizon ($t > 8$; task loss at $t = 8$), so task supervision cannot see it. **Left:** accuracy vs. step; without a penalty the late lean erodes accuracy after the horizon; the drift penalty restores the calibrated trajectory; the symmetric penalty flattens the gain as well. **Right:** residual lean $|\beta|$ by objective.

counting. Tokens charged = trigger position plus 48 tokens per forced-answer branch executed by the policy (a fixed allocation, not measured decoded length); fixed stops and stop-policies are charged one branch per executed forcing, majority voting is charged all eight, and full budget is charged none, including when its terminal answer required forcing. Critic evaluation and branch prefill are not charged. These are archived token allocations, not measured wall-clock time or FLOPs. *(vii) Onset null.* Onset localization is reported against a within-chain permutation null (random step position), with grid resolution 512 tokens noted. *(viii) Placebo normalization and failure criterion.* The placebo is assessed as the predictable fraction $\mathbb{E}[\hat{m}^2]/\,\mathbb{E}[\Delta^2]$ against a probe-matched shuffled control at comparable increment variance. If the verdict probe's normalized predictable fraction exceeds $3\times$ its shuffled control while its final-correctness AUC exceeds 0.75, we will report this as evidence against the tapped-state harmonicity approximation, not as probe estimation error. *(ix) Interpretation rules.* If $\mathrm{AUC}(-\hat{B})$ is statistically indistinguishable from the supervised classifier, probe-slope, or length-only baselines (paired bootstrap), we will report that mid-chain states generically predict failure and that the compensator's contribution is semantic (interpretation, onset localization, the readout–verdict gap) rather than raw detection power. If the drift alarm does not Pareto-dominate checkpoint-majority voting and verdict-probe stopping under accounting (vi), we will report the alarm as one member of a family of effective stopping rules rather than as superior. *(x) Resampling study.* If compute allows, $k = 4$ additional chains are sampled for $\sim$100 level-3–5 problems to measure label stability under resampling, the per-problem accuracy-vs-budget estimand, a sibling-chain shared-drift test, and a token-matched self-consistency point for the Pareto curve.

Preliminary AIME-2024 audit (post-hoc). As a hardest-regime probe we ran $k = 4$ chains per problem on all 30 AIME-2024 problems at 1.5B with the identical protocol (120 chains; final accuracy 17.5%; 88% hit the 4096-token cap). No chain is overthink-wrong (0 of 120) and the capped-cohort curve rises monotonically ($0.9\% \to 10.4\%$, $n = 106$), but this configuration is overwhelmingly competence-limited: 99 chains are never correct, and only 12 of 120 chains ever reach a correct early checkpoint and are therefore *eligible* for an overthink-wrong transition at all. On that opportunity set the count is 0/12, whose 95% upper bound is $\approx$25%. The zero is a low-powered null, not evidence that overthinking is absent where folklore expects it; budgets beyond 4096 remain untested.

Forcing-operator validation (post-hoc). To test whether the budget-forced answer matches a *natural* halt, we truncated 160 stratified (chain, checkpoint) pairs at 1.5B, closed the think block *without* the forcing string, and let the model continue greedily for up to 220 tokens. Most natural continuations (69.4%) emitted no boxed answer at all—absent the forcing string the model tends to resume reasoning rather than commit, which is why the operator is needed in the first place. Where a natural answer *did* appear, it matched the forced answer string in 81.6% of cases and agreed on correctness in 85.7%; the residual disagreement is consistent with genuine indeterminacy of

mid-chain commitment under two different halting prompts. We therefore read $\hat{\pi}_t$ as commitment *under the standard budget-forcing operator* (the object our conventions define), with agreement measured only among the minority of prefixes that answer under both protocols; this does not bound unconditional operator error.

Checkpoint-grid robustness (post-hoc). Recomputing labels on a coarser 1024-token grid (a strict subset of recorded forcings): overthink-wrong counts move $10 \to 9$ at 1.5B and $9 \to 3$ at 7B. Because the grids are nested and the overthink-wrong label requires a correct answer at *some* checkpoint, refining the grid can add chains but never remove one: the count is non-decreasing in density, so the $\approx 3\%$ figure is specific to the 512-token grid—a ceiling for coarser grids and a floor for finer ones, not a one-sided bound.

The k=4 quota resample is not a prevalence estimate. We do *not* pool the k=4 quota-selected resample set (item x) into a prevalence estimate; the pooled estimate is the separate k=8 study below. That set was selected by outcome-dependent quotas on the original chain's label, so it is a census of the 10 overthink-wrong problems and is enriched $\approx 4.6\times$ for the outcome being estimated; that design supports label stability, not prevalence. Post-stratifying by the known population strata gives 3.2% $[1.0\%, 6.4\%]$ for a fresh chain (against 4.7% unweighted on the enriched sample), which merely recovers the independent single-chain rate $10/367$ and so adds little information—the interval is wide because the overthink-wrong stratum contains only 10 problems.

Resampling study (pre-registered item x; results). $k = 4$ fresh chains per problem on 80 level-3–5 problems (1.5B, identical protocol). Label stability: 3/10 overthink-wrong problems reproduce the label in at least one sibling; 38/40 stay-correct problems remain stay-correct. The sibling-chain shared-drift test finds no problem-level adverse-drift signal (the original chain's mid-chain erosion does not predict sibling overthinking; Pearson $r = -0.21$), and self-consistency over the five chains (76.2% at 15771 tokens) does not beat one full-budget chain here. This quota sample is too small to settle whether flips are chain-idiosyncratic; the larger k=8 study below, on the complete problem set, finds a clear problem-level propensity (23.8% of overthink-wrong siblings against 2.1% of stay-correct siblings) alongside substantial chain noise, and supersedes this reading.

k=8 resampling (pre-registered). Eight fresh chains per problem, from eight disjoint seed streams, on all 367 level-3–5 problems at 1.5B, under the audit protocol otherwise unchanged (T=0.6, top-p 0.95, 4096-token cap, 512-token forcing grid, identical labels): 2936 chains, of which 100 overthink-wrong, 1989 stay-correct, 249 late-bloomer and 598 early-unresolved/wrong. Unlike item (x) the problem set is complete rather than quota-selected, so pooling is legitimate: the rate 3.4% (problem-clustered 95% CI $[2.4\%, 4.5\%]$) is the prevalence quoted in the abstract and Sec. 5.3; Table 1 lists the single-chain counts. As the dated plan specifies, it supersedes the single-chain $10/367$ as the headline. Sibling propensity and the non-monotonicity screen are reported in Sec. 5.3.

Exp 0 details. Evidence separation $d = 0.35$, $\sigma = 1$, $T = 60$ ($\varepsilon = 0.06$ in Fig. 10); the drift arm adds ε to each log-odds increment (probability-scale drift $\approx \varepsilon\pi(1-\pi)$; slope/$R^2$ are measured against the exact $b(h,t)$ by quadrature, not raw ε). Overthinking variant: initial evidence scale 1.0 decaying as $e^{-t/8}$ with the per-step Bayes gain, lean persistent. Objective panel: $T = 40$, learnable (gain, lean), fixed representational noise $\tau = 0.12$ (hence the learned gain ≈ 0.46 above the noise-free Bayes gain 0.35), scarce-label task loss (96 labeled problems), penalty on unlabeled rollouts, $\lambda \in \{0, 0.3, 1, 3, 10, 30\}$, critic 5 inner steps per reasoner step.

H.1 Post-hoc robustness analyses (not pre-registered)

The analyses in this subsection were run *after* unblinding, on the frozen rollout and state caches. They are sensitivity checks on the pre-registered conclusions, not new pre-registered claims.

Survivorship decomposition: empirical terms (Prop. 5). Conditioning follows the published curve's convention ($R(T)$: checkpoint T recorded, $= \{\tau \geq T\}$ on the grid, matching Prop. 5; capped chains included at the cap); the reproduced aggregate curve matches the published one to machine precision and the identity is exact per step. Region totals over $2048 \to 4096$ (chain-level

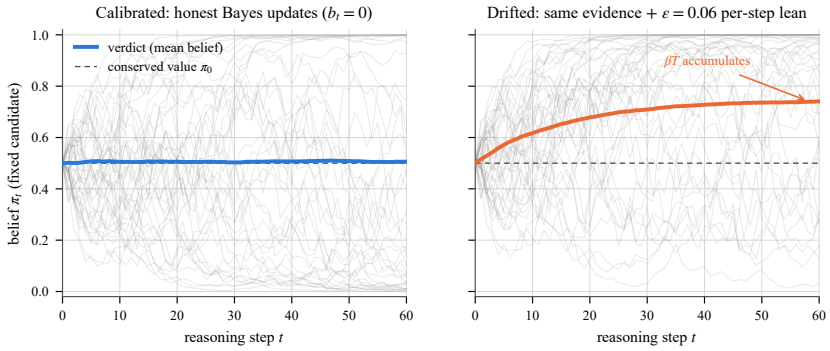


Figure 10: Bayesian evidence accumulation (belief = posterior on a fixed candidate). **Left:** honest updates ($b_t = 0$): chains diffuse to $0/1$, the mean belief (= verdict) is conserved (max deviation 0.010). **Right:** the same evidence with a constant lean: the mean bends toward a confidently wrong value, ballistically.

bootstrap, 2000 resamples): 1.5B marginal change -8.3 = within $+8.5$ $[+2.1, +14.5]$ + composition -16.8 $[-22.3, -12.0]$; 7B $-4.4 = +20.0$ $[+12.6, +27.9] + -24.4$ $[-31.9, -17.6]$ points. The within-cohort term is significantly *positive* at both scales—chains still reasoning keep improving—so survivorship more than explains the entire marginal decline.

	1.5B				7B			
T	n	$M(T)\%$	within	comp	n	$M(T)\%$	within	comp
512	367	30.0	–	–	262	25.6	–	–
1024	358	50.6	+21.5	-0.9	257	44.7	+20.2	-1.1
1536	319	56.1	+11.0	-5.4	236	58.1	+16.9	-3.6
2048	270	56.7	+6.7	-6.1	197	50.8	-2.5	-4.8
2560	222	51.8	+3.6	-8.5	166	51.8	+7.2	-6.2
3072	192	50.0	+2.1	-3.9	146	51.4	+2.7	-3.2
3584	172	50.6	+3.5	-2.9	125	53.6	+8.0	-5.8
4096	151	48.3	-0.7	-1.6	97	46.4	+2.1	-9.3

Label-definition robustness. Relaxing the strict pre-registered overthink-wrong definition: dropping the $\leq$ 2048-token requirement gives 13/367 (3.5%) at 1.5B and 9/262 (3.4%) at 7B—at 7B identical to the base count: every early-correct-then-wrong chain was already correct by token 2048. Requiring two consecutive correct checkpoints gives 8 (2.2%) and 3 (1.1%). Counting chains with *any* correct→wrong forced-answer flip anywhere gives 40 (10.9%) and 26 (9.9%)—but most flipped chains recover by the natural end. Prevalence stays within ∼1–4% under every final-outcome definition.

Cumulative-history-summary ablation. This follow-up was specified before its own results, but after the original audit. Fixed-window taps can forget information: a pair tap at $t + 1$ no longer explicitly includes h_{t-1}. Moreover, `multilayer_tap.py` enriches only the critic in its pair variant, leaving both probes on single states. We therefore refit both probes and both critics with $z_t = [h_t, \text{mean}(h_0, \ldots, h_t)]$ at layer 21. The splits, seed, calibration fraction, hyperparameters, and shuffled-label diagnostic are unchanged. A single-state control reproduces 0.325/0.418 and 0.300/0.359.

This construction is history-dependent but not necessarily nested. For example, histories $(h_0, h_1, h_2) = (0, 1, 0)$ and $(1, 0, 0)$ give different z_1 but identical $z_2 = (0, 1/3)$. Hence the construction does not guarantee $\sigma(z_t) \subseteq \sigma(z_{t+1})$. The filtration $\sigma(z_0, \ldots, z_t)$ is nested; conditioning on z_t alone is a different operation. A compressed recurrent state likewise requires a sufficiency argument before inheriting the verdict martingale.

The verdict diagnostic decreases from 0.418 to 0.362 at 1.5B and from 0.359 to 0.336 at 7B, while the commitment diagnostic also falls (0.325 to 0.295; 0.300 to 0.255). The planned pos-

itive commitment-minus-verdict gap is not obtained (-0.067/-0.081). These results show only that this augmentation does not resolve the control discrepancy; they cannot distinguish observation loss from estimation error. Nor does the tower property require this ordering for fitted, variance-normalized diagnostics.

There are additional confounds. The doubled input dimension raises the 7B commitment shuffled-label diagnostic from 0.063 to 0.106 (42% of the measured value), limiting comparisons across feature sets. The implementation computes running means over checkpoint states for probe fitting but over paragraph-step states for the increment series, so the history-summary sampling rules differ between training and application. A conclusive history control would require consistent observation sequences, an explicitly nested conditioning history, and separate assessment of estimation error. The archived `nested_tap` filenames and numerical macros retain their original names for traceability, not as a claim of nested information.

Critic capacity. Replacing the MLP critic with a single linear layer (identical protocol and splits): at 1.5B the linear critic retains genuine held-out skill on increment prediction ($R^2 = 0.16$ vs. the MLP's 0.26), so the predictable-drift signal is not an expressivity artifact of the MLP. The energy-ratio metric is not meaningful for the linear critic (its shuffled-label null inflates), and at 7B the 3585-dimensional linear map overfits its $\sim$4.3k training rows ($R^2 = -1.29$; high estimator variance)—the 7B half of this ablation is uninformative, and the 1.5B half carries the conclusion.

State-richness series under the variational score. Commitment and verdict variance-normalized scores as the tap is enriched (identical protocol and splits; *single* = pre-registered layer, $+$*history* = one-step hidden-state history, *3 layers* = L14/21/27 concatenation), reported as $\widehat{V}/\widehat{\mathrm{Var}}(\Delta)$ rather than $\widehat{\mathbb{E}}[g^2]/\widehat{\mathrm{Var}}(\Delta)$, with the score difference from a *paired* chain-level bootstrap (one shared resample index for both readouts, 2000 resamples):

	1.5B		7B	
tap	c/v	gap [95% CI]	c/v	gap [95% CI]
single layer	0.284/0.249	$+0.036$ [+0.007,+0.063]	0.221/0.244	-0.023 [-0.050,+0.007]
+ history	0.232/0.231	$+0.001$ [-0.025,+0.025]	0.172/0.227	-0.055 [-0.082,-0.027]
3 layers	0.278/0.296	-0.018 [-0.056,+0.021]	0.290/0.292	-0.002 [-0.035,+0.032]
3 layers + history	0.299/0.285	$+0.014$ [-0.023,+0.052]	0.016/-2.096	$+2.112$ [+1.832,+2.467]

A positive commitment-minus-verdict gap is what an idealized, accurately estimated verdict control would suggest; the tower property alone does not imply this ordering for fitted, variance-normalized scores. The gap is significant only at the 1.5B single-layer tap; elsewhere the gap straddles zero or, at the 7B one-step-history tap, has a CI excluding zero in the reversed direction (one of eight post-hoc cells, uncorrected). The 7B three-layer-plus-history row is degenerate: its critic diverges at $21{,}505$ input dimensions and yields a strongly negative variational value. A labeled post-hoc rerun at a smaller learning rate gives gap $+0.015$ $[-0.020, +0.052]$. Bootstrap intervals are conditional on the fitted models; they do not include probe or critic refit uncertainty, and enrichment need not improve a finite-sample estimate.

Randomized refits are diagnostics, not calibrated hypothesis tests. The archived analysis used 200 refits per scheme at the primary taps and 50 at richer taps: a global training-step shuffle and a circular shift of the *concatenated* training target vector. The latter is not a within-chain shift. Refits used at most 80 epochs, versus 200 for the observed critic. Neither scheme establishes exchangeability for the heterogeneous, nonstationary chains, so its exceedance ranks are not calibrated permutation p-values; we make no significance claim from them. The released implementation also provides a separately versioned within-chain shift with matched fit settings; that correction has not been run for the results reported here. Even within-chain shifts require additional stationarity or exchangeability assumptions for inferential use. The positive held-out estimates and their chain-bootstrap intervals, rather than these ranks, support the reported probe-increment predictability.

The older $\mathbb{E}[g^2]$ statistic has a positive estimation-noise floor: its ratio to the shuffled control varies across refits. The population variational identity avoids that particular inflation, but does not make the empirical estimate a guaranteed bound or require every randomized-refit value to be negative. All archived numeric diagnostics are retained in the result files.

Base-family audits: R1-Distill-Llama 8B and Phi-4-mini-reasoning (endpoints fixed in the dated plan before either run). The Qwen3 audit varies training recipe but not base lineage, so we repeated the audit on two reasoning models with non-Qwen bases: DeepSeek-R1-Distill-Llama-8B (DeepSeek-AI et al., 2025), on a Llama-3.1 base, in 4-bit NF4 as in the 7B escalation; and the 3.8B Phi-4-mini-reasoning (Xu et al., 2025), on a Phi-4 base, in fp16. Both models emit an explicit thinking block, which the forcing operator requires. The pipeline is otherwise unmodified—one chain per problem on the same 367 level-3–5 MATH-500 problems, temperature 0.6, 4096-token cap, the same 512-token forcing grid and labels. Neither audit is floored or ceilinged (final accuracy 78.7% and 82.6%). Labels: 8 overthink-wrong (5 strict), 252 stay-correct, 37 late-bloomer, 70 early-unresolved/wrong for Llama; 8 (5), 272, 31, 56 for Phi. Both headline findings reproduce outside the Qwen lineage. Prevalence is 2.2% [1.1%, 4.2%] and 2.2% [1.1%, 4.2%], just below the ≈3% of the two primary Qwen-based R1-distills. The marginal decline over $2048 \to 4096$ (-9.3 and -7.5 points) again decomposes into a *positive* within-cohort term ($+16.6$ [$+9.3$, $+24.1$]; $+9.1$ [$+1.8$, $+16.6$]) against composition (-25.9 [-31.8, -20.9]; -16.6 [-22.7, -11.1]). Deviations are recorded in `results_llama8b/` and `results_phi4mini/`: the generation batch was reduced from the pipeline default because a larger batch exceeded the 12 GB card—on this platform the driver spills to host memory rather than raising an error, so the reduction was triggered by a throughput floor—and each run was interrupted and resumed through the pipeline's own resume logic (on (problem, checkpoint) for forcing, which decodes greedily and is therefore deterministic; rollout resumption re-seeds as the Qwen3 run did). Every record was generated exactly once. These runs measure the two pre-registered endpoints only: they extract no hidden states, so no probe, critic or stopping replay is fitted for them and they do not appear in Table 2. They are two additional families, not a scaling study.

Third-model audit: Qwen3-1.7B (endpoints fixed in the dated plan before the run; post-hoc relative to the original audit). Qwen3 shares a base family with our R1-distills but is independently trained and not R1-distilled, so this varies training recipe and distillation status, not base lineage; we do not read it as evidence of cross-family generality. The full pipeline—one chain per problem on the same 367 level-3–5 MATH-500 problems, temperature 0.6, 4096-token cap, the same 512-token forcing grid and labels—applied to Qwen3-1.7B (Yang et al., 2025). The audit is informative rather than floored or ceilinged (final accuracy 78.2%, 63% at 2048). Labels: 11 overthink-wrong (5 strict), 253 stay-correct, 34 late-bloomer, 69 early-unresolved/wrong. Both headline findings reproduce: prevalence 3.0% [1.7%, 5.3%] brackets the ≈3% of the two R1-distills, and the marginal decline over $2048 \to 4096$ (-2.6 points) again decomposes into a *positive* within-cohort term ($+7.6$ [$+1.6$, $+13.7$]) against composition (-10.2 [-14.3, -6.1]). Deviations, recorded in `results_qwen3/qwen3_audit.json`: generation batch 6 rather than 8 (VRAM headroom; throughput only), and the rollout was interrupted by external process kills three times and resumed via the pipeline's resume logic, so the sampling stream is not bit-identical to one uninterrupted seed-0 run—all 367 chains are complete, deduplicated, and generated once.

A forcing-free view: length–accuracy curves are censoring-dominated (post-hoc). The decomposition above reads forced answers. As a check that the forcing operator is not manufacturing the result, we repeat the exercise with no forcing in the analysis: bin each chain by its natural termination length L and score it by its *own* unforced final answer. The familiar declining curve appears, steeply (-9.1 points per 512-token bin at 1.5B, -8.6 at 7B). Two findings follow, and the first does *not* match the mechanism our forced decomposition isolates. (i) Stratifying by MATH level removes little of the decline—pooled within-level slopes are -8.4 / -8.4, so difficulty composition accounts for only ≈8% / ≈3% of it. (ii) It is almost entirely *censoring*: restricting to chains that actually terminated ($L < 4096$) collapses the slope to -2.4 [-4.5, -0.6] and -1.1 [-2.8, $+0.3$]. The long-length bins are populated by chains that hit the cap without ever committing to an answer (41% / 37% of chains). Correspondingly, the surviving pool degrades by -20.2 / -23.7 points over $2048 \to 4096$ under this natural-answer scoring—a same-direction, larger effect than the forced $M(T)$ moves, though a different estimand (it carries no "what would it answer mid-chain" content). *Caveat:* for capped chains the "final" answer must itself be elicited by forcing, so this analysis is strictly forcing-free only on the terminated subset—which is exactly where the effect is smallest. Reading: aggregate accuracy-vs-length plots are dominated by *which chains occupy which bin*—here through censoring, in Fig. 2 through termination-induced composition. Both are selection; neither is within-chain erosion. The forced decomposition remains the stronger design, since it conditions on chain identity and needs no difficulty proxy at all.

Decoding-matched forcing (pre-registered). Chains are sampled (T=0.6, top-p 0.95) but forced answers are decoded greedily, so the readout is taken under a different rule than the process it describes. We re-forced every checkpoint of the 240 chains above under the chain's own decoding rule, 3 draws each, and compared paired on identical (chain, checkpoint) pairs. The difference is negligible: greedy exceeds sampled forced-answer correctness by $+0.7$ points, 95% problem-clustered CI $[-0.3, +1.8]$ over 120 problems, with per-checkpoint deltas between -0.3 and $+2.4$ points. overthink-wrong counts are 5 under greedy and $6/7/4$ across the three sampled draws. We recorded in advance the prediction that greedy forcing would *inflate* the overthink-wrong count, making the reported prevalence an upper bound in that respect; **that prediction is not borne out**—the sampled counts are if anything marginally higher, and the spread across draws exceeds the greedy-sampled difference. The defensible reading is that the operator's decoding rule does not materially affect the measurement, not that either rule is conservative.

Dense-grid test (pre-registered). We recorded in advance the prediction that a *finer* forcing grid weakly increases the overthink-wrong count, because each extra checkpoint is another chance to catch a transient correct answer—the same mechanism that made the coarse-grid recount lower the count. Re-forcing the T=0.6 chains of the previous paragraph on a 256-token grid (a strict superset of the 512 and 1024 grids) confirms it: 5 overthink-wrong chains at 1024 tokens, 5 at 512, and 7 of 240 at 256. The crosstab isolates the mechanism exactly: all 5 chains labeled overthink-wrong on the 512 grid keep the label, and the increase comes entirely from 2 chains that the coarser grid had recorded as *never-correct*—the finer grid finds a brief correct answer between 512-token checkpoints that is subsequently lost. Prevalence measured this way is therefore non-decreasing in checkpoint density (unchanged from 1024 to 512, up at 256), so the 512-grid figure is a floor relative to finer nested grids and a ceiling relative to coarser ones—grid-specific rather than conservative.

Decoding-temperature ablation (pre-registered). On a stratified 120-problem subset with k=2 chains per problem, we repeat the audit at the pipeline default (T=0.6) and at T=0.9, everything else identical. Prevalence does not rise with temperature: 3 of 240 chains are overthink-wrong at 0.9 (1.2%, [0.4%, 3.6%]) against 5 of 240 at 0.6 (2.1%, [0.9%, 4.8%])—if anything lower, though the intervals overlap and this subset is small. The structural finding reproduces at both settings: the marginal curve peaks near 1.5–2k tokens and declines to the cap ($56\% \to 37\%$ at 0.6; $60\% \to 48\%$ at 0.9) while the final-step within-cohort delta stays non-negative—$+1.1$ points $[-4.2, +6.4]$ at 0.6 and $+8.4$ $[+3.3, +14.8]$ at 0.9, the latter significantly positive. Per the pre-registration this is a sensitivity report, not a gate.

Probe time-input ablation (post-hoc). The probes receive the tapped state only, while the critic additionally receives normalized token position, which raises the question whether the verdict probe's residual predictable movement is an omitted-time artifact. The tested time-input augmentations do not explain it away. Normalized position is linearly decodable from the tap itself ($R^2 = 0.75$–0.92 out-of-sample), so the probes were never time-blind; refitting both probes with position added (and with a scaled polynomial in position, to rule out the feature being crushed by the shared penalty) moves the verdict fraction by $+0.011/+0.004$ at 1.5B and ≤ 0.001 at 7B, all smaller than the 0.014–0.037 standard deviation of the critic refit itself; and a critic given *time alone* reaches a fraction of only 0.001–0.006. Probe AUC/ECE reproduce the published values exactly in the no-time arm.

Head-to-head with published stopping rules (post-hoc). We replay named methods over the recorded chains under a single effective-token accounting: budget forcing / fixed truncation (Muennighoff et al., 2025), self-consistency (Wang et al., 2023), Adaptive-Consistency (Aggarwal et al., 2023), ESC (Li et al., 2024), First-Finish Search (Agarwal et al., 2025), answer-consistency (Liu & Wang, 2025), Dynasor-CoT (Fu et al., 2024), ES-CoT (Mao et al., 2025), our alarm, and an oracle bound. At 1.5B each method is labeled *pure replay* (28/48), *approximate replay* (16, with the deviation recorded per row), or *not implementable* (4); the 7B replay has 28 rows, 21 of them implementable. Confidence- and entropy-gated exits (e.g. Kadavath et al., 2022) fall in the last class: we cached hidden states and answer strings but not token logits, so replaying them would require new generation, and we do not approximate them. Two step-level rules (ES-CoT, answer-consistency) never fire at their published hyperparameters because our 512-token forcing grid is $\approx 10\times$ coarser than a reasoning step; this is a limitation of our cache, not evidence against those methods, and

grid-rescaled variants are reported as *our* rescaling. Outcome: 0 methods beat the full budget at lower cost; the accuracy gains that survive bootstrap are all multi-sample and cost 3–5× the tokens. Against a token-matched fixed truncation our alarm is within $+0.3$ points—i.e. at equal compute it is not better than truncating, which we regard as the sharpest negative result about our own controller.

Self-consistency inherits the survivorship structure. Majority-vote and best-of-N budget curves over the resampled MATH chains (k=4 plus the original) and AIME chains (k=4), with endpoints pre-registered in the supplement: the vote curve never declines within cohorts beyond noise anywhere on the grid, and over $2048 \rightarrow 4096$ its pooled within-cohort term is $+12.0$ points $[+4.6, +20.2]$ against composition -12.0 $[-17.7, -7.3]$ (problem-level bootstrap, 2000 resamples); best-of-N shows the same structure ($+5.4$ within, -9.3 composition). Thus changing-cohort selection also affects vote curves; this does not establish shared adverse drift or exclude benefits from voting at other operating points. Caveats: 80+30 problems give wide intervals, and the AIME split is nearly degenerate (one problem leaves the survivor set by the cap).

Alarm-threshold sensitivity. Sweeping the alarm's calibration percentile over $\{50, 60, 70, 75, 80, 85, 90, 95, 99\}$ with everything else fixed: the accuracy-vs-tokens trade-off is monotone at both scales and no threshold improves accuracy over the full budget. The sweep's nominal best (99th percentile: $+0.0$ points at 1.5B, $+0.0$ at 7B) is degenerate—it ties the full budget by (almost) never firing—so the pre-registered 90th-percentile operating point is not beaten by any non-trivial alternative, and the absence of an observed accuracy-improving alarm operating point persists across this threshold sweep.

Synthetic mechanism validation across regimes. One-factor-at-a-time around the Exp 0 base configuration (drift strength, innovation noise, horizon, drift onset, critic training-sample size; 19 grid points × 10 seeds): Doob compensation recovers a median 98.6% of injected accuracy loss (range 36.9–99.5%), robust down to 500 critic-training problems (Fig. 11). The one failure mode is low innovation noise ($\sigma = 0.5$, recovery 36.9%): beliefs saturate near $0/1$, where probability-scale drift vanishes and the compensator cannot flip wrong-side saturated paths—the same saturation pathology that defeats symmetric magnitude penalties (Sec. 5.1), here appearing as a limit of compensation itself.

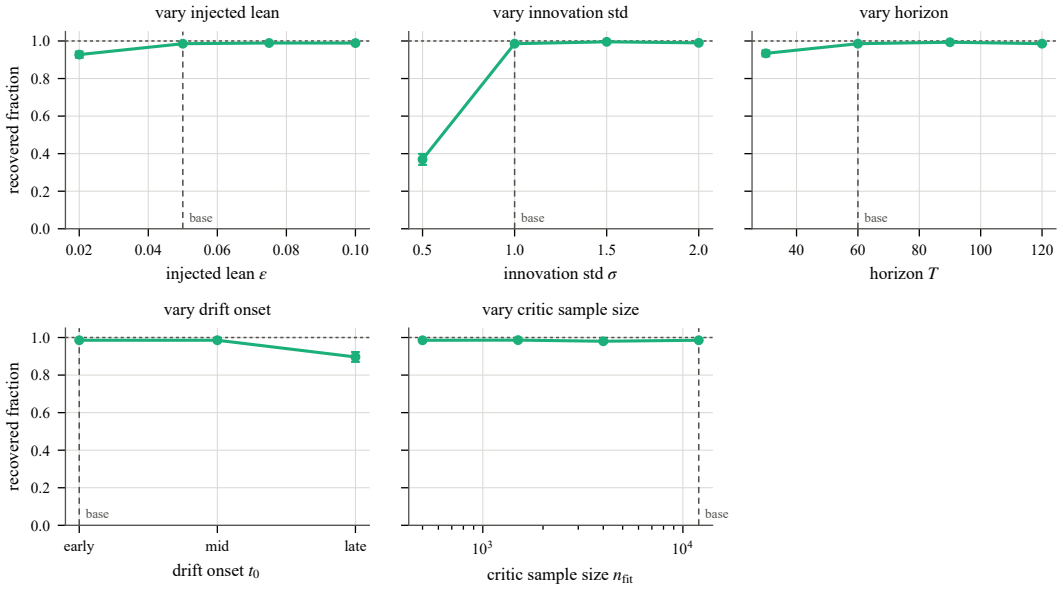


Figure 11: Recovered fraction of injected accuracy loss under Doob compensation, varying one factor at a time around the Exp 0 base configuration (marked); error bars are s.e. over 10 seeds. Recovery is high across all factors except very low innovation noise, where beliefs saturate.

I Observation Histories, Loss Hazards, and Stopping Gain

These consequences of conditional expectation clarify what the measured conservation property can mean; they are not proposed as new projection or optional-stopping theorems.

Proposition 19 (An observation-calibrated verdict can drift)**.** *Let $Y \in [0,1]$, $\mathcal{G}_t \subseteq \mathcal{F}_t$, and $v_t = \mathbb{E}[Y \mid \mathcal{G}_t]$. The observation fields $\mathcal{G}_t$ need not be nested. Write $\mathsf{P}_t X = \mathbb{E}[X \mid \mathcal{G}_t]$ and $e_t = v_t - \Phi_t$, where $\Phi_t = \mathbb{E}[Y \mid \mathcal{F}_t]$. Then*

$$m_t^{\mathcal{G}} := \mathbb{E}[v_{t+1} - v_t \mid \mathcal{G}_t] = (\mathsf{P}_t\mathsf{P}_{t+1} - \mathsf{P}_t)Y = \mathbb{E}[e_{t+1} \mid \mathcal{G}_t], \qquad \|m_t^{\mathcal{G}}\|_2 \le \|e_{t+1}\|_2.$$

If $\mathcal{G}_t \subseteq \mathcal{G}_{t+1}$, the conditional drift is zero. Without nesting, exact observation-level calibration alone does not imply this conclusion.

Proof. The first identity substitutes the two conditional expectations and uses $\mathsf{P}_t^2 = \mathsf{P}_t$. For the second, $\mathbb{E}[\Phi_{t+1} \mid \mathcal{G}_t] = \mathbb{E}[Y \mid \mathcal{G}_t] = v_t$ by $\mathcal{G}_t \subseteq \mathcal{F}_t \subseteq \mathcal{F}_{t+1}$. Conditional expectation contracts L^2 norms. Nesting gives the final claim by the tower property. □

In particular, $\mathcal{H}_t = \sigma(z_0, \dots, z_t)$ is a filtration and $\mathbb{E}[Y \mid \mathcal{H}_t]$ is a bounded martingale, even if $\mathcal{H}_t \subsetneq \mathcal{F}_t$. Retained observation history is therefore another sufficient route to an exact verdict null; full-state sufficiency is not necessary. This does not make a learned history probe exact, nor does it identify why the real-model control is nonzero.

Exact four-path control. Draw $Y \sim \mathrm{Bernoulli}(1/2)$ and a signal S with $\Pr(S = Y \mid Y) = 4/5$. At $t = 0$ the observation is S; at $t = 1$ it is constant; at $t = 2$ it is Y. The full history retains S throughout. Conditional on $S = 1$, the full verdict is $(4/5, 4/5, Y)$ and the compressed verdict is $(4/5, 1/2, Y)$; for $S = 0$ they are $(1/5, 1/5, Y)$ and $(1/5, 1/2, Y)$. Every compressed value is exactly $\mathbb{E}[Y \mid z_t, t]$. Its first conditional increment is $-3/10$ or $+3/10$, while its second has zero conditional mean given the current constant observation. Averaged over the two steps, predictable squared energy is $9/200$, compared with zero for the full verdict. Retaining S at $t = 1$ restores the null. All four path probabilities $(2/5, 1/10, 1/10, 2/5)$ are enumerated with exact rational arithmetic. This is a possibility result about measurement, not evidence that forgetting causes the LLM verdict residual.

Proposition 20 (Binary commitments and loss hazards)**.** *Let $A_t \in \{0,1\}$ be the exact correctness of the fixed greedy stopping operator, adapted to $\mathcal{F}_t$. Set $p_{10,t} = \Pr(A_{t+1} = 0 \mid \mathcal{F}_t)$ and $p_{01,t} = \Pr(A_{t+1} = 1 \mid \mathcal{F}_t)$. Then*

$$b_t = (1 - A_t)p_{01,t} - A_t p_{10,t}, \qquad (b_t^-)^2 = A_t p_{10,t}^2, \qquad b_t^2 = A_t p_{10,t}^2 + (1 - A_t)p_{01,t}^2.$$

Proof. Condition on $A_t = 0$ or $A_t = 1$ in $b_t = \mathbb{E}[A_{t+1} - A_t \mid \mathcal{F}_t]$. Only the latter case permits negative drift; squaring gives the two identities. □

Thus the one-sided objective penalizes squared loss hazard at correct states, whereas the symmetric objective also penalizes repair at incorrect states. A binary martingale is constant almost surely; a submartingale can repair errors while making the correct set absorbing. These statements do not supply an observable correctness oracle or a tested training method.

Corollary 21 (Adverse energy bounds stopping gain)**.** *Under Cor. 4, define $\mathcal{E}_- = \mathbb{E}[\sum_{t<\tau}(b_t^-)^2]$. For any admissible $S \le \tau$,*

$$\mathbb{E}[C_S - Y] \le \min\{1, \sqrt{\mathbb{E}[\tau]\mathcal{E}_-}\}.$$

Proof. The stopping identity bounds the gain by $\mathbb{E}[\sum_{t<\tau} b_t^-]$. Cauchy–Schwarz on the product measure of paths and pre-termination time indices bounds this by $\sqrt{\mathbb{E}[\tau]\,\mathbb{E}[\sum_{t<\tau}(b_t^-)^2]}$. Its total measure is $\mathbb{E}[\tau] < \infty$; $|b_t| \le 1$ ensures integrability. The gain is also at most one. □

For binary commitments, $\mathbb{E}[\sum_{t<\tau} b_t^-]$ equals the expected number of correct-to-wrong transitions. The reported early-correct/final-wrong frequency is not this count: it omits later onset and losses subsequently repaired, so it cannot be inserted into this bound.

J PROTOCOL-MATCHED CONTINUATION FOLLOW-UP

Design and selection. This post-hoc follow-up was specified in an internal plan after the primary results were known, with the sample fixed before collecting the new outcomes. All 12 historical wrong-ending 1.5B chains with a correct nonterminal checkpoint at or before 2048 are included, across all five MATH levels. For each first-correct checkpoint, one unique right-ending chain with a correct answer at that same checkpoint is selected, with an exact difficulty match. A seeded shuffle followed by minimum level-distance matching gives the 12 pairs in Table 4; no new outcome enters selection. All 24 immediate greedy answers reproduce the archived correct label.

Protocol fidelity and runtime. The cached R1-Distill-Qwen-1.5B model is run in FP16 on the same 12 GB RTX 4070 Ti. Prompt tokens are reconstructed with the original chat template; continuations start from the original generated prefix token IDs, without decoding and retokenizing that prefix. At checkpoint t, each of eight continuations has $4096 - t$ new-token allowance, temperature 0.6, and top-p 0.95, ending at EOS or the cap. Batches contain four draws; a deterministic per-prefix/per-batch seed derived from 20260907 permits exact resumption within the recorded runtime. The frozen manifest records input and source hashes, model revision, sampling configuration, and selected IDs. Runtime is PyTorch 2.13.0+cu126 and Transformers 5.15.1; cross-version bitwise reproduction of historical generation is not assumed.

Immediate forcing and terminal forcing reproduce the original text-based operator: close the thought if necessary, append the boxed-answer cue, and decode greedily for at most 48 tokens. At termination, a boxed answer after the closing thought is used if present; otherwise the forcing operator is applied. Forced outputs use the original brace-closing parser and symbolic verifier. A strictly balanced-box parser is also evaluated and changes no correctness label among the 192 continuations. No continuation lacks a scored answer. Terminal forcing is used in $85/96$ case and $28/96$ control draws; respectively $11/96$ and $68/96$ reach EOS. Thus the fixed-budget protocol still involves substantial censoring, but unlike the historical branching diagnostic it does not score a missing unforced answer as an automatic error. Raw tokens, answer strings, scoring flags, prompt/generation token counts, and timings are retained. Generation of the continuations and their terminal forcing uses 405,150 tokens including EOS, excluding the immediate preflight checks; this is not a matched FLOP or wall-clock policy comparison.

Estimates and uncertainty. For arm $a \in \{\text{case}, \text{control}\}$ and pair $i = 1, \ldots, 12$, let Y_{aij} be terminal correctness of continuation $j = 1, \ldots, 8$. The fixed immediate answer is correct at every selected prefix, so

$$\widehat{\Phi}_{ai} = \frac{1}{8}\sum_{j=1}^{8} Y_{aij}, \qquad \widehat{d}_{ai} = 1 - \widehat{\Phi}_{ai}, \qquad \widehat{D} = \frac{1}{12}\sum_{i=1}^{12}(\widehat{d}_{\text{case},i} - \widehat{d}_{\text{control},i}).$$

Thus prefixes have equal weight; the 192 draws are not 192 independent problems. The contrast is descriptive of these selected cohorts, not a randomized treatment effect. Each arm has 12 prefixes and 96 new continuations. Immediate correctness is one at every prefix; continuation success is $48/96$ in cases and $94/96$ in controls. The equal-prefix mean gaps are 50.0 (95% prefix-bootstrap interval $[28.1, 71.9]$) and 2.1 ($[0.0, 6.2]$) percentage points. The case-minus-control contrast is 47.9 points ($[27.1, 69.8]$). The 5000-resample paired bootstrap draws entire matched pairs and recomputes the two means; it summarizes pair-level sampling variation rather than providing an exact population interval. Per-prefix Wilson intervals in Table 4 expose the limited precision of eight Monte Carlo draws. Conditional plug-in Monte Carlo standard errors of the two arm means are 3.43 and 1.36 percentage points, but can understate uncertainty at all-success/all-failure prefixes; the Wilson intervals should not be replaced by zero uncertainty there.

Specifically, NumPy's default random generator with seed 20260907 draws 5000 vectors of 12 pair indices with replacement. The same indices are used for both arms, and the reported limits are the 2.5th and 97.5th percentiles. Draws are not independently resampled inside prefixes. These descriptive intervals therefore do not separately propagate all within-prefix Monte Carlo uncertainty or confer coverage for an outcome-independent population. For an arm with success estimates $\widehat{\Phi}_i$, the reported conditional plug-in Monte Carlo standard error is $\sqrt{\sum_{i=1}^{12} \widehat{\Phi}_i(1 - \widehat{\Phi}_i)/7/12}$. The per-prefix Wilson intervals use eight Bernoulli trials and $z = \Phi_{\text{N}}^{-1}(0.975)$, with Φ_{N} denoting the

Table 4: All matched pairs. IDs index MATH-500; the first ID is the historical wrong-ending case and the second its right-ending control. Every immediate answer is correct. Brackets are per-prefix 95% Wilson intervals for continuation success, not intervals for population prevalence.

Case/control	Level	Checkpoint	Case correct	Interval	Control correct	Interval
9/109	5	2048	1/8	$[0.02, 0.47]$	8/8	$[0.68, 1.00]$
101/63	5	1536	6/8	$[0.41, 0.93]$	8/8	$[0.68, 1.00]$
120/57	2	1536	7/8	$[0.53, 0.98]$	8/8	$[0.68, 1.00]$
126/461	4	1024	2/8	$[0.07, 0.59]$	6/8	$[0.41, 0.93]$
128/471	3	512	0/8	$[0.00, 0.32]$	8/8	$[0.68, 1.00]$
238/459	5	1536	7/8	$[0.53, 0.98]$	8/8	$[0.68, 1.00]$
274/328	5	2048	8/8	$[0.68, 1.00]$	8/8	$[0.68, 1.00]$
326/35	2	2048	1/8	$[0.02, 0.47]$	8/8	$[0.68, 1.00]$
381/313	5	512	0/8	$[0.00, 0.32]$	8/8	$[0.68, 1.00]$
382/232	5	2048	2/8	$[0.07, 0.59]$	8/8	$[0.68, 1.00]$
398/73	4	512	6/8	$[0.41, 0.93]$	8/8	$[0.68, 1.00]$
493/323	5	1536	8/8	$[0.68, 1.00]$	8/8	$[0.68, 1.00]$

standard normal distribution function. These summaries quantify different sources of variation and are not interchangeable.

Scope of the inference. At a specified prefix, the matched success gap estimates expected remaining adverse drift by Eq. (2). It does not identify the mechanism of a particular flip. Because the cohorts were selected using archived final outcomes, this follow-up does not revise the 3.4% prevalence, estimate loss at an outcome-independent sample of prefixes, or validate a deployable stopping policy. Difficulty and checkpoint matching do not remove all differences between problems. The historical cross-protocol diagnostic in App. H is preserved separately; its mean continuation success is not pooled with this experiment.

K Whole-Problem Refit Sensitivity

This post-hoc diagnostic tests sensitivity to the fitted measurement pipeline, not a new behavioral intervention. Each model uses its original layer-21 states and problem-disjoint outer probe/critic/evaluation split. Logistic regularization is fixed at $C = 0.1$, rather than retuned by the original cross-validation grid. Within the probe role, the original 20% calibration problems remain disjoint from probe fitting. A fixed 15% of eligible critic-role problems form a disjoint validation set, replacing the original step-level validation partition. Ten seeds, 20260907–20260916, independently resample whole problems with replacement within each development role and final-label stratum, then refit both probes and both critics. The evaluation problems never enter fitting, calibration, early stopping, or model selection.

Probe fitting uses logistic regression (maximum 2000 iterations) and temperature calibration. The critic retains the two-layer 256-unit MLP, dropout 0.1, AdamW with learning rate 10^{-3} and weight decay 10^{-4}, batch size 1024, maximum 200 epochs, and validation-NLL early stopping with patience 12. Dense paragraph increments and normalized token position follow the primary pipeline. Every fitted pipeline is evaluated on the same held-out problems using $\widehat{V}/\widehat{\mathrm{Var}}(\Delta)$; the two readouts use identical evaluation steps. Complete development resampling lists, probe AUC/ECE/Brier scores, temperatures, critic validation diagnostics, and per-step held-out predictions and targets are saved.

Commitment exceeds verdict in $4/10$ and $1/10$ refits, respectively. The mean gaps are -0.088 and -0.029; Table 5 retains all fits, including the negative 1.5B commitment score in refit 2. That probe has temperature 101.02, ECE 0.020, and increment variance 2.22×10^{-4}, while its critic predicts mean increment 0.0137 against observed mean 0.00017 and performs worse than the zero predictor in the variational score. This is an observed sensitivity, not a reason to discard the run or evidence for a unique cause. A small ECE does not guarantee drift estimation quality. With ten development refits and fixed evaluation problems, these results are not a reliable 95% interval for full-pipeline

Table 5: All ten refits of the diagnostic variant. C/V denote commitment/verdict scores $\widehat{V}/\widehat{\mathrm{Var}}(\Delta)$; gap is C minus V. Ranges across these refits are sensitivity summaries, not confidence intervals.

	1.5B			7B		
Refit	C	V	Gap	C	V	Gap
1	0.2201	0.2332	-0.0130	0.2254	0.2342	-0.0088
2	-0.6384	0.2512	-0.8896	0.1837	0.2427	-0.0590
3	0.2372	0.2553	-0.0180	0.2403	0.2161	0.0242
4	0.3045	0.2790	0.0255	0.2193	0.2546	-0.0353
5	0.2449	0.2539	-0.0090	0.2161	0.2437	-0.0276
6	0.2194	0.1964	0.0230	0.1857	0.1862	-0.0005
7	0.2702	0.2460	0.0242	0.1804	0.2774	-0.0970
8	0.1921	0.2532	-0.0611	0.1857	0.2227	-0.0370
9	0.2999	0.1953	0.1047	0.2368	0.2458	-0.0090
10	0.1708	0.2347	-0.0639	0.2095	0.2484	-0.0389

uncertainty. They qualify the original single-fit ordering without replacing its numerical results or resolving the real-model verdict null.

L Supplement Guide and Reproduction Boundaries

The unified `supplementary_code.zip` contains the primary audit and the separately versioned follow-up, together with archived manuscript sources and audit figures. The present MAIN2 source and `main2_related.bib` are separate manuscript files; App. M adds no experimental results. The root README gives an artifact index and commands. The original result files and numerical macros are retained; new measurements are isolated under `results/research_20260907/` and `paper/followup_numbers.tex`. Internal plans record when designs were fixed, not external registration or independent replication.

Claim or appendix	Evidence and verification route
Primary audits; App. H	Original `results/`, `results7b/`, `results_qwen3/`, `results_llama8b/` and `results_phi4mini/` JSONs; `code/make_numbers.py` regenerates the historical macro file.
Observation control; App. I	`code/research_exact_control.py` enumerates the four paths with rational arithmetic; `exact_control.json` stores the 29 checks.
Matched continuations; App. J	`matched_manifest.json`, `matched_runtime.json`, `matched_immediate.jsonl`, `matched_draws.jsonl`, and `matched_summary.json`; all are in the follow-up results directory.
Refits; App. K	`refit_config.json`, role lists, 20 individual JSON/NPZ output pairs, and `refit_summary.json`; NPZ files contain held-out targets, critic predictions, and problem IDs.

Model-free verification. The archive provides a standard-library SHA-256 integrity check. From a fresh extraction, run:

```
python code/verify_supplement.py
```

Adding `--evidence` requires NumPy and recomputes the matched success rates, paired-bootstrap intervals, per-prefix Wilson intervals, token totals, parser-label agreement, and all saved refit scores. It also checks frozen source hashes, problem-role separation, numerical follow-up macros, and the exact finite-state construction. Checks run in a temporary copy and leave the packaged evidence unchanged. They verify arithmetic and consistency, not the validity of the original answer verifier or independent replication of model behavior.

Additional assets and computation requirements. The archived 1.5B rollout and forced-answer files needed for the matched manifest are included. Model weights, dataset caches, and the large primary hidden-state arrays are not. New continuations require the recorded model revision, dataset,

generation runtime, and GPU resources. Repeating the refits requires the original layer-21 state arrays and index files for both models; saved held-out predictions suffice only for score recomputation, not refitting. The historical README documents the extraction pipeline. Different runtime versions can change generated trajectories and therefore all downstream fitted quantities. Independent reruns must use a new output location or version, preserving the frozen evidence.

Evidence that remains absent. The matched experiment does not sample prefixes independently of archived outcomes, and the refit diagnostic is neither an original-tuning replication nor a full-pipeline confidence interval. Neither experiment supplies a held-out deployment benefit or identifies the cause of the verdict residual. The historical cross-protocol branching study and randomized-refit ranks remain archived diagnostics with their original limitations; they are not silently replaced by the new follow-up.

M Beyond Single-Chain Reasoning: Scope and Testable Extensions

The central distinction is between forecasting a fixed protocol's eventual success and evaluating the artifact that could be delivered now. The unit of computation need not be a token or a single-model reasoning segment. This section makes that extension explicit, without claiming empirical validation outside the settings in Sec. 5.

A protocol-level formulation. Fix a finite horizon T and a complete stochastic protocol, including its history-dependent choices of models, tools, messages and updates. Let $\mathcal{H}_t$ be its nested full-history filtration, $C_t \in [0,1]$ the score obtained by a specified deliver-now operator, and $u_t = \mathbb{E}[C_t \mid \mathcal{H}_t]$ its current commitment. Fix the scoring rule and evaluation distribution. Let $Y \in [0,1]$ be the terminal score and require a common terminal boundary $u_T = Y$, with the terminal evaluation included in $\mathcal{H}_T$. Define

$$v_t = \mathbb{E}[Y \mid \mathcal{H}_t], \qquad d_t = \mathbb{E}[u_{t+1} - u_t \mid \mathcal{H}_t].$$

The tower property and conditional telescoping give

$$\mathbb{E}[v_{t+1} \mid \mathcal{H}_t] = v_t, \qquad u_t - v_t = -\,\mathbb{E}\left[\sum_{s=t}^{T-1} d_s \,\middle|\, \mathcal{H}_t\right]. \tag{8}$$

For a history-measurable stopping time $S \le T$, the same argument yields $\mathbb{E}[C_S - Y] = -\,\mathbb{E}[\sum_{s=S}^{T-1} d_s]$. Early termination can be represented by freezing the score thereafter. These are applications of the existing identities to a bounded score and a joint protocol, not new mechanisms or guarantees of improvement. They need no independence among agents or Markov assumption; a state-based Poisson equation additionally requires a sufficient Markov state. If the stopping readout and terminal score do not satisfy the common-boundary condition, a terminal mismatch term must be retained.

Multi-step reasoning and iterative self-revision. For a single task, let the current artifact be the answer after each reasoning segment or completed critique–revision round. Self-Refine, for example, iteratively generates feedback and revises an output (Madaan et al., 2023). The commitment is the current artifact's externally scored quality; the verdict forecasts the final artifact under the remaining refinement protocol. More elaborate feedback, increased confidence, or a changed answer does not alone identify their difference. A direct audit samples prefixes without selecting on their eventual outcomes, scores the immediate artifacts, and branches fresh continuations under the same protocol. This estimates remaining improvement or erosion at those states, while repeated runs quantify sampling variability. The theory permits beneficial revision and does not prescribe zero drift: positive drift is precisely the improvement that refinement seeks.

Multi-agent collaboration and debate. Multiagent debate uses several model instances to exchange and revise reasoning before producing a common answer (Du et al., 2023). Here $\mathcal{H}_t$ contains the joint interaction history, including relevant private states and tool observations, and the deliver-now operator specifies how the current messages or candidate answers become one system

output. Equation (8) concerns that system output, not the average confidence or agreement of its members. A transcript-only probe may be lossy; even an exact forecast conditioned on one agent's local information need not be a martingale relative to the larger joint filtration. Thus the information set must be specified before interpreting predictable updates as a control failure.

Communication can change both dependence and expected answer quality. The independent-ensemble $1/\sqrt{k}$ innovation scaling discussed in App. D therefore does not automatically extend to interacting agents. Neither consensus nor correlated errors alone determine the sign of remaining drift. A useful audit fixes the final-answer aggregation rule, measures quality and agreement separately at each round, and resamples continuations from the same joint state. Comparisons among communication protocols require their own verdicts and compute accounting. If only unfinished problem instances remain in a later-round average, the survivor decomposition applies to that changing problem cohort; replacing an agent inside a fixed problem is not itself this population-selection effect.

Recursive improvement of the system itself. Improving an answer is distinct from modifying the machinery that generates answers. The Darwin Gödel Machine, for example, modifies agent code and evaluates the resulting agents on coding benchmarks (Zhang et al., 2025c). To formulate an analogous audit, the state must include the evolving code, parameters, memory, selection history and update procedure. The current artifact is now the system version that would be deployed if development stopped, and commitment measures its performance under a fixed evaluation distribution. A fixed *meta-protocol* may contain adaptive, even self-modifying, updates: "fixed" specifies the overall probability law, not frozen model parameters.

Under that specification and the common terminal boundary, Eq. (8) still concerns the forecast of one terminal evaluation. It neither fixes the performance of intermediate versions nor rules out recursive capability improvement. Externally changing the update procedure or evaluation target changes the comparison; a verdict for the old protocol cannot simply be carried over. A frozen probe also need not retain its meaning when the underlying representations change (App. G). A credible empirical extension would separately evaluate successive deployable versions on held-out tasks, account for selection and development compute, and use replicated update trajectories where feasible. Our present experiments establish none of these system-improvement claims.

Implication for LLM development. Across these settings, the framework supplies a common comparison target: the externally evaluated artifact available now versus the artifact expected after additional computation. It separates measurement from mechanism and decision quality. This can inform how future evaluations, process supervision and adaptive-compute methods are designed, but it does not establish that our present estimators or stopping heuristics transfer successfully to these systems.