Calibration-Free Early Stopping for Chain-of-Thought Reasoning via E-processes

Dong-Jae Lee* **Gyojin Han*** **Junmo Kim**
KAIST
{jhtwosun, hangj0820, junmo.kim}@kaist.ac.kr

Abstract

Reasoning large language models (LLMs) benefit from test-time scaling, motivating adaptive policies that spend tokens when they help and stop early otherwise. Existing rules halt when a confidence or stability signal crosses a threshold, but such thresholds rarely transfer across models and datasets, and validation-calibrated rules lose their guarantee under shifts in test-time horizon or signal dynamics. We reframe adaptive stopping as sequential evidence accumulation along the reasoning trajectory rather than threshold calibration on held-out data. Our framework runs two e-processes on a shared per-chunk signal: a commit stream tracks the signal's position relative to its slow trend, and an abstain stream tracks its signed velocity, making the two anti-correlated by construction. Under signal-level nulls on the per-chunk increments, Ville's inequality yields an anytime-valid signal-level false-alarm bound at user-specified significance levels, with no held-out calibration. Across various model–dataset combinations, the framework matches or exceeds calibration-based stopping on the token–accuracy Pareto front using zero calibration data, and the same significance level transfers unchanged across cells where calibration becomes infeasible.

1 Introduction

Reasoning large language models (LLMs) (Yang et al., 2025; Guo et al., 2025) benefit from test-time scaling: allocating more reasoning tokens often improves performance on mathematical, scientific, and symbolic tasks (Snell et al., 2025; Brown et al., 2024). Long chain-of-thought generation has accordingly become a standard inference-time tool, but it also creates a compute-allocation problem. Some instances are solved early and continue to consume tokens after the answer has effectively stabilized; others exhaust the full budget without making useful progress. The central challenge is to make models reason adaptively: spending computation when it improves reliability, and stopping when further reasoning is unlikely to change the answer.

Recent adaptive-reasoning methods (Yang et al., 2026; Fu et al., 2026; Aggarwal et al., 2023; Li et al., 2024) instantiate this idea by monitoring a confidence or stability signal during generation and halting once the signal crosses a threshold. This converts a fixed-budget policy into an instance-dependent stopping rule, but only by trading budget selection for threshold selection: thresholds on confidence, entropy, or stability rarely transfer across models and datasets. Risk-control approaches such as Conformal Thinking (CT) (Wang et al., 2026b) address this by calibrating thresholds to a user-specified error rate on a held-out validation set. The guarantee, however, is conditional on the calibration distribution matching deployment: adaptive reasoning runs across many combinations of models, benchmarks, and prompt formats, where per-cell calibration is often unavailable or infeasible at strict risk targets under modest validation budgets. A source-calibrated rule continues to run when transferred to a new cell, but its guarantee no longer applies.

The difficulty is structural: a chain-of-thought trace is not a single prediction but a sequence of dependent intermediate states, each offering another opportunity to stop. The event to control is therefore whether an unsafe stopping rule ever fires along the trajectory, not whether a single held-out example crosses a threshold. Calibrating before deployment ties the rule to one cell's statistics. When the test-time horizon, signal dynamics, or instance mix shifts, the calibrated mapping no longer matches the trajectory at hand. A stopping rule must therefore handle the sequential structure of the trajectory without depending on calibration data that may be unavailable or fail to transfer.

*Equal contribution.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 35798–35814
October 24-29, 2026 ©2026 Association for Computational Linguistics

We address both requirements by introducing e-processes (Ramdas et al., 2023; Shafer, 2021) to the chain-of-thought stopping problem: sequential evidence accumulators whose error guarantees hold at every stopping time, with no held-out calibration. We instantiate them as CES (Calibration-Free Early Stopping), a sequential-evidence framework for adaptive chain-of-thought stopping. Rather than testing a raw signal against a calibrated threshold, CES accumulates evidence along the reasoning trajectory and stops the first time this evidence is sufficient to either commit to or rule out the current answer. The strength of evidence is governed by a user-controlled significance level α: under a signal-level null on the per-chunk increment, Ville's inequality (Ville, 1939) bounds the false-alarm probability over the entire trajectory by α at any stopping time. Because this bound is signal-level, the rule requires no held-out calibration, and α retains its meaning across models, datasets, and token budgets.

We further propose a dual construction with two e-processes operating on a shared signal s_t, which can be any of the trajectory statistics already explored in adaptive-reasoning work (e.g., confidence, entropy, answer stability). The two streams target complementary decisions: the commit stream tracks the position of s_t relative to its own slow trend, and the abstain stream tracks its signed velocity. They are anti-correlated by construction: an ascending s_t feeds commit evidence and depletes abstain evidence in the same step, so they rarely fire together. Together, the streams form a symmetric exit rule on the shared signal, so the stopping decision does not depend on any single mapping from signal to correctness.

Our contributions are: (i) We introduce CES, a sequential-evidence framework for adaptive chain-of-thought stopping that provides an anytime-valid signal-level guarantee via Ville's inequality, with a single user-controlled confidence parameter α and no held-out calibration data. (ii) We propose a dual e-process construction on a shared signal that handles commit and abstain symmetrically under a single decision rule, using a position-velocity decomposition that makes the two streams anti-correlated by construction. (iii) We show that CES matches or exceeds Conformal Thinking's token–accuracy Pareto across various model–dataset combinations using zero calibration data, and applies unchanged under cross-cell transfer where CT's calibration becomes infeasible at strict risk targets or fails to transfer.

2 Related Work

Adaptive stopping for LLM reasoning. Recent work modulates test-time compute by monitoring an intermediate signal: token-level entropy or top-1 probability (Zhang et al., 2025; Yang et al., 2026), answer-token entropy at intermediate predictions (Wang et al., 2026a), intermediate-answer self-consistency (Wang et al., 2023; Aggarwal et al., 2023; Li et al., 2024), learned probes (Ni et al., 2026), or budget-aware policies (Snell et al., 2025; Fu et al., 2026). These methods either pre-set a threshold heuristically or calibrate it on a held-out set.

Conformal prediction for LLMs. Conformal prediction methods derive split-conformal guarantees from a held-out calibration set (Angelopoulos and Bates, 2023; Yadkori et al., 2024). Conformal Thinking (Wang et al., 2026b) adapts split conformal to the reasoning-trace setting with a dual upper/lower stopping rule controlled by a UCB-Hoeffding margin. The deployed bound is a fixed-time empirical guarantee tied to the calibration distribution; the bound does not transfer to a domain with shifted trace length or signal dynamics.

E-processes and anytime-valid inference. E-processes (Ramdas et al., 2023; Shafer, 2021) are sequential evidence accumulators built from a betting metaphor: a fictitious gambler stakes a fraction of current wealth against the null hypothesis at each step. Under H_0, the gambler cannot profit in expectation, so large accumulated wealth constitutes evidence against H_0. Ville's inequality (Ville, 1939) formalizes this: for any nonnegative supermartingale, the probability of the wealth *ever* exceeding $1/\alpha$ is at most α, at any stopping time (Ramdas et al., 2023; Shafer, 2021). This time-uniformity distinguishes e-processes from classical p-values, which require a fixed sample size and break under adaptive stopping. The betting fraction controls the tradeoff between power and conservatism. Various predictable strategies exist, ranging from fixed values to adaptive schemes such as ONS and aGRAPA (Waudby-Smith and Ramdas, 2024) that maximize expected log-wealth growth without prior knowledge of the signal distribution. We instantiate this framework for chain-of-thought stopping by defining a chunk-level filtration and mapping trajectory signals to bounded e-value incre-

ments via an EMA-deviation transform. For more formal background, please refer to Appendix B.

3 Method

3.1 Setup

We consider a reasoning LLM that produces a chain-of-thought trace $r_{1:T} = (r_1, \ldots, r_T)$ of T chunks for a query x with gold answer y. After observing $r_{1:t}$, the model yields a running prediction $f_t(x)$ and an auxiliary scalar $s_t \in \mathbb{R}$ that summarizes the per-step signal associated with f_t. The scalar s_t may be an answer confidence, the stability of f_t across recent steps, or any other primary signal used by an early-exit policy. Prior work focuses on designing better primary signals (Fu et al., 2026; Wang et al., 2026a); we instead study the early-exit policy that consumes them.

An early-exit policy selects a stopping time $\tau \leq T$ and takes one of two actions: (i) *commit*, returning $f_\tau(x)$ when further reasoning is judged unnecessary; or (ii) *abstain*, deferring to a fallback when the instance appears unsolvable within the full budget. Both actions reduce token usage relative to running to T while intermediate probing costs are explicitly accounted for in our evaluation.

3.2 Goal

Reasoning horizons and signal dynamics vary across models, datasets, and token budgets, so any rule tuned on a held-out set is unlikely to transfer. Models also encounter instances beyond their capacity, far from the validation distribution, where additional reasoning consumes tokens without improving the answer. The optimal threshold likewise depends on these factors, so no fixed value reliably trades accuracy for efficiency.

We therefore design an early-stopping rule with three properties: (i) no dependence on a held-out calibration set; (ii) symmetric handling of commit and abstain under a single decision rule; and (iii) user-specified control over the significance level α.

3.3 E-Process Construction

Instead of asking whether a raw signal s_t has crossed a calibrated threshold, we test whether the trajectory has accumulated sequential evidence in favor of stopping. To this end, we construct dual e-processes: E_t^+ accumulates evidence for commit, and E_t^- accumulates evidence for abstain. Crossing the upper boundary indicates that further reasoning is redundant; crossing the lower boundary indicates that further reasoning is unlikely to improve the answer and triggers the fallback policy.

Each stream maps s_t to a bounded evidence increment $\phi_t^m \in [-1, 1]$ for $m \in \{+, -\}$ and updates a nonnegative wealth process

$$E_t^m = \prod_{i=1}^{t} \big(1 + \lambda_i^m \phi_i^m\big), \tag{1}$$

with $E_0^m = 1$. The betting fraction $\lambda_i^m \geq 0$ is predictable—it depends only on $\mathcal{F}_{i-1}$—and satisfies $1 + \lambda_i^m \phi_i^m \geq 0$ so that E_t^m remains nonnegative. In practice, λ_i^m can be fixed, selected from a small portfolio, or set adaptively (Section 3.5).

We define a signal-level null hypothesis corresponding to the absence of systematic positive evidence drift by requiring the evidence increment to have nonpositive conditional mean,

$$H_0^m : \quad \mathbb{E}\big[\phi_t^m \,\big|\, \mathcal{F}_{t-1}\big] \leq 0. \tag{2}$$

Since λ_t^m is $\mathcal{F}_{t-1}$-measurable, taking conditional expectation in Eq. 1 under H_0^m gives

$$\begin{aligned} \mathbb{E}\big[E_t^m \,\big|\, \mathcal{F}_{t-1}\big] &= E_{t-1}^m \big(1 + \lambda_t^m \, \mathbb{E}[\phi_t^m \mid \mathcal{F}_{t-1}]\big) \\ &\leq E_{t-1}^m, \end{aligned} \tag{3}$$

so E_t^m is a nonnegative supermartingale starting at 1 under the null.

For any such process, Ville's inequality (Ramdas et al., 2023) bounds the probability that it ever exceeds a level $\alpha_m > 0$:

$$\Pr_{H_0^m}\left(\sup_t E_t^m \geq 1/\alpha_m\right) \leq \alpha_m. \tag{4}$$

Unlike Markov's inequality, the supremum is taken over the entire trajectory rather than at a fixed t, so the bound holds at *every stopping time*, including those chosen adaptively from the current data. This anytime validity allows us to monitor E_t^m continuously without inflating the false alarm rate. Then, we define the per-stream stopping time

$$\tau_m = \inf\{t : E_t^m \geq 1/\alpha_m\}. \tag{5}$$

Stopping at τ_m therefore raises a false alarm with probability at most α_m, uniformly in t.

3.4 Position-Velocity Decomposition

What remains is to choose the signal s_t and its mapping to the bounded evidence ϕ_t^m. Rather than designing a separate signal for each stream, we use a single s_t and extract two complementary statistics: its position relative to a slow trend, and its velocity.

Commit increment (position). Let

$$e_t = \gamma s_t + (1-\gamma)\, e_{t-1}, \qquad 0 < \gamma \ll 1, \quad (6)$$

denote a slow exponential moving average (EMA) of s_t. The commit increment is the clipped deviation from this trend,

$$\phi_t^+ = \mathrm{clip}(s_t - e_t,\, -1,\, 1). \quad (7)$$

The commit stream accumulates evidence when s_t exceeds its slow trend, removing the need for an absolute threshold on s_t, a quantity whose scale shifts across models and datasets.

Abstain increment (velocity). The abstain increment is the signed velocity of s_t,

$$\phi_t^- = \mathrm{sign}(s_{t-1} - s_t), \quad (8)$$

positive when s_t declines, negative when it rises, and zero otherwise. The abstain stream therefore accumulates evidence when the trajectory loses stability.

The two increments are anti-correlated by construction: ϕ_t^+ is non-decreasing in s_t, whereas ϕ_t^- is non-increasing in s_t; consequently, their bet-weighted increments are conditionally anti-correlated.

3.5 Betting Rate

Another design choice is the betting fraction λ_t^m. Two requirements constrain it: λ_t^m must be predictable so that Ville's bound applies, and it should adapt to the realized drift of ϕ_t^m—unknown a priori—without any held-out tuning. We adopt the aGRAPA (Waudby-Smith and Ramdas, 2024),

$$\lambda_t^m = \mathrm{clip}\left(\frac{\sum_{s<t} \phi_s^m}{1 + \sum_{s<t} (\phi_s^m)^2},\, 0,\, \lambda_{\max} \right), \quad (9)$$

with $\lambda_{\max} = 1/2$ for $1 + \lambda_t^m \phi_t^m \geq 1/2 > 0$. The rate depends only on $\phi_{<t}^m$, so it is $\mathcal{F}_{t-1}$-measurable. Under strong positive drift the numerator $\sum_{s<t} \phi_s^m$ grows, λ_t^m saturates near $\lambda_{\max}$, and E_t^m accelerates, while under weak drift λ_t^m stays near zero and evidence accrues slowly.

3.6 Dual E-Process Early Exits

Combining the per-stream stopping times τ_+ and τ_- with the budget T, the overall exit time is

$$\tau = \min\{\tau_+,\, \tau_-,\, T\}. \quad (10)$$

The action depends on which boundary triggers first: the policy commits to $f_\tau(x)$ if $\tau = \tau_+$, abstains if $\tau = \tau_-$, and returns the full-budget answer $f_T(x)$ if neither boundary is crossed before T.

3.7 Sensitivity Control and Guarantee

The significance levels α_+, α_- are user-specified parameters that control the commit and abstain thresholds. Each α_m acts as a monotone sensitivity knob. A smaller α_+ raises the threshold $1/\alpha_+$, delays τ_+ on average, and yields fewer commits, more tokens, and (empirically) a lower task-level false-positive rate; the same monotonicity holds for α_- on the abstain stream. The pair (α_+, α_-) therefore controls the commit/abstain trade-off without calibration data; the formal guarantee is signal-level (see Appendix C), and its empirical task-level behavior is reported in Section 4.4.

Two invariances make the α-based control portable. The bound is anytime-valid: it holds uniformly over the entire trajectory, so the rule can inspect E_t^m at every chunk and stop at the first crossing without inflating the error rate. The bound is also distribution-free, depending only on user-selected confidence level α_m; combined with the local-trend normalization of ϕ_t^m, the same α governs decisions across models, datasets, and token budgets without calibration data.

4 Experiments

4.1 Experimental Setup

Models and datasets. We apply our method to three reasoning LLMs: Qwen3-8B, Qwen3-30B-A3B-Instruct-2507 (Yang et al., 2025), and DeepSeek-R1-Distill-Qwen-32B (Guo et al., 2025). We evaluate across four mathematical reasoning benchmarks: AIME, MATH-500 (Lightman et al., 2024), OlympiadBench (He et al., 2024), and DeepScaler (Luo et al., 2025). AIME comprises 1,011 competition problems drawn from the American Invitational Mathematics Examination spanning 1983 to 2024. MATH-500 is a 500-problem subset of the MATH benchmark covering a broad range of high-school competition topics. OlympiadBench is a multilingual olympiad-level benchmark, from which we use the English math subset of 674 problems. DeepScaler is a curated collection of 1,189 challenging math problems designed to stress-test advanced reasoning models. Token use includes the reasoning prefix together with every forcing-suffix prefill and intermediate-answer decode performed up to the stopping boundary. CT-EAT and CT-Stab incur the same probes as their CES counterparts, while truncation and confidence-only methods do not require answer probes. Position-aware FLOPs accounting and detailed overhead statistics are re-

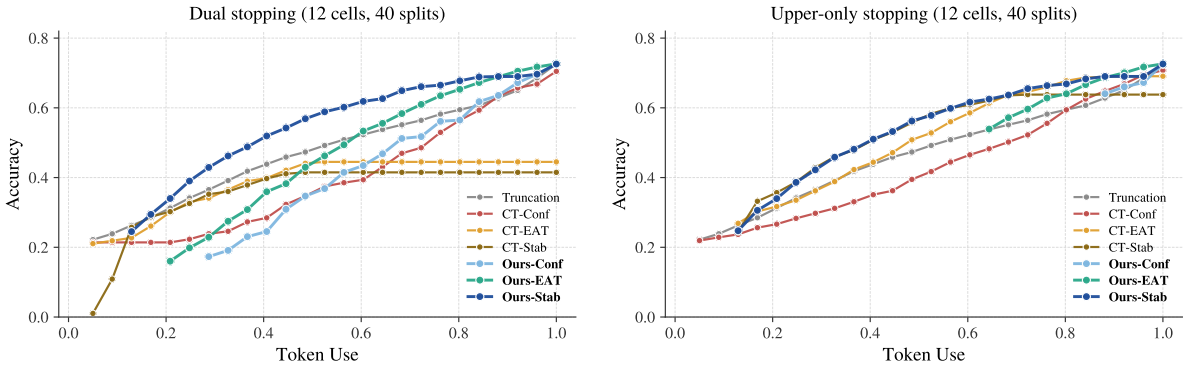


Figure 1: Token-use vs. accuracy Pareto curves aggregated over all 12 model–dataset combinations. The left panel shows dual stopping, where early commit and early abstain are both active so the model can either commit to an intermediate answer or defer to a fallback policy before the trajectory ends. The right panel shows upper-only stopping, where only early commit is applied and the model falls back to the terminal answer otherwise.

	Dual stopping						Upper-only stopping					
TokU	**CT-Conf**	**CT-EAT**	**CT-Stab**	**Ours-Conf**	**Ours-EAT**	**Ours-Stab**	**CT-Conf**	**CT-EAT**	**CT-Stab**	**Ours-Conf**	**Ours-EAT**	**Ours-Stab**
≤ 0.2	0.214	0.286	0.302	0.131	0.158	**0.337**	0.275	0.286	0.310	—	—	**0.317**
≤ 0.3	0.243	0.347	0.352	0.180	0.235	**0.441**	0.310	0.364	**0.408**	—	—	**0.408**
≤ 0.4	0.280	0.398	0.397	0.245	0.358	**0.517**	0.363	0.453	0.485	—	—	**0.491**
≤ 0.5	0.357	0.441	0.415	0.364	0.444	**0.574**	0.408	0.545	0.538	—	—	**0.567**
≤ 0.6	0.394	0.445	0.415	0.434	0.530	**0.616**	0.502	0.595	0.614	—	0.508	**0.637**
≤ 0.7	0.486	0.445	0.415	0.512	0.594	**0.656**	0.557	0.628	0.631	—	0.591	**0.637**
≤ 0.8	0.562	0.445	0.415	0.564	0.653	**0.677**	0.589	0.683	0.641	0.611	0.636	**0.690**
≤ 0.9	0.645	0.445	0.415	0.662	**0.697**	0.690	0.654	0.683	0.641	0.670	**0.700**	0.690
≤ 1.0	0.705	0.445	0.415	**0.726**	**0.726**	**0.726**	0.707	0.683	0.641	**0.726**	**0.726**	**0.726**

Table 1: Accuracy at each token budget, averaged over 40 splits and 12 cells. CT methods use $n_{\text{val}} = 50$ while Ours uses zero calibration data. CT-Stab applies CT's calibration rule to the same stability signal used by Ours. —: no Pareto operating point exists at or below that budget (the e-process commit threshold was not crossed at any α in the sweep). **Bold**: highest value in each half-row.

ported in the supplementary material.

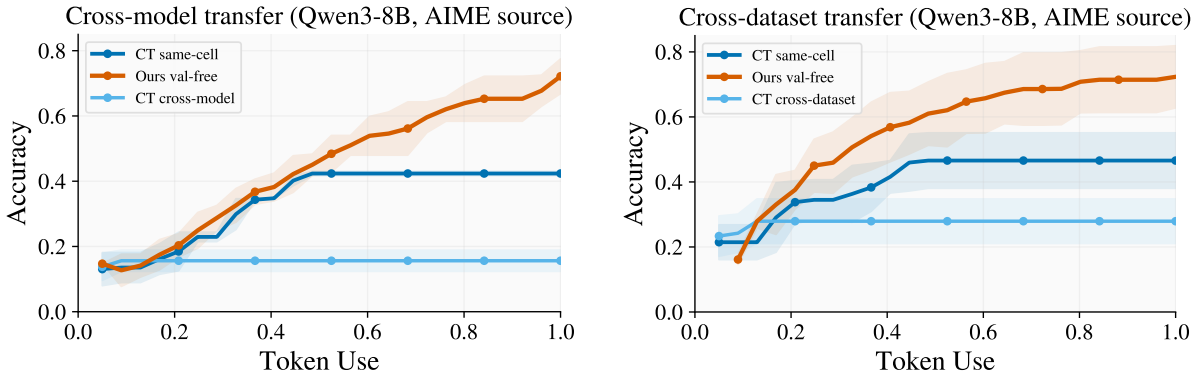
Trace collection. Each problem is presented to the model with the instruction to reason step by step, separating reasoning steps by blank lines, and to write the final answer inside `\boxed{}`. Trajectories are generated using vLLM (Kwon et al., 2023) following the official sampling parameters recommended by each model's authors; for Qwen3 models we use T=0.6, p=0.95, k=20 (with T=0.7, p=0.8 for the Qwen3-30B-A3B variant), and for DeepSeek-R1 we use T=0.6, p=0.95. The output text is segmented into reasoning chunks at blank-line boundaries.

Intermediate predictions and signals. At each chunk boundary, we append a forcing suffix (`\n\n**Final Answer**\n\boxed{`) to the partial trace and complete greedily (T=0, $\leq$24 tokens) to extract an intermediate boxed prediction $\hat{y}_t$. Three scalar signals are derived from the partial trace at each boundary: Confidence (Yang et al., 2026) (geometric mean of per-token probabilities), EAT (Wang et al., 2026a) (inverted first-answer-token entropy), and Stability (normalized run length of consecutive equal predictions). Each signal is smoothed with a causal EMA (α=0.3) before entering the e-process; the bet-construction EMA uses γ=0.05. Formal definitions are in Appendix A.

Protocol. We follow CT's (Wang et al., 2026b) evaluation protocol: 40 random splits, validation size $n_{\text{val}} = 50$, test capped at 800. For ours, the validation set is unused at deployment; for CT, it is used to UCB-calibrate the upper threshold τ (and the lower sigmoid steepness c in the dual variant). Sweeps cover $\alpha \in \{10^{-4}, \ldots, 0.95\}$ for ours and target risk $\varepsilon \in \{0.05, 0.10, \ldots, 1.00\}$ for CT.

4.2 Experiment Results

We evaluate the proposed method against CT and a truncation baseline across all 12 model–dataset combinations (3 models $\times$ 4 datasets), reporting



(a) Cross-model: different models on AIME.

(b) Cross-dataset: Qwen3-8B on other datasets.

Figure 2: Token-use vs. accuracy Pareto under cross-combination deployment with Qwen3-8B/AIME as the fixed calibration source. CT same-cell and the proposed method are evaluated on the same target cells as reference envelopes. CT cross-cell performance degrades substantially when the calibration distribution does not match the target cell, whereas the proposed method operates identically without re-calibration.

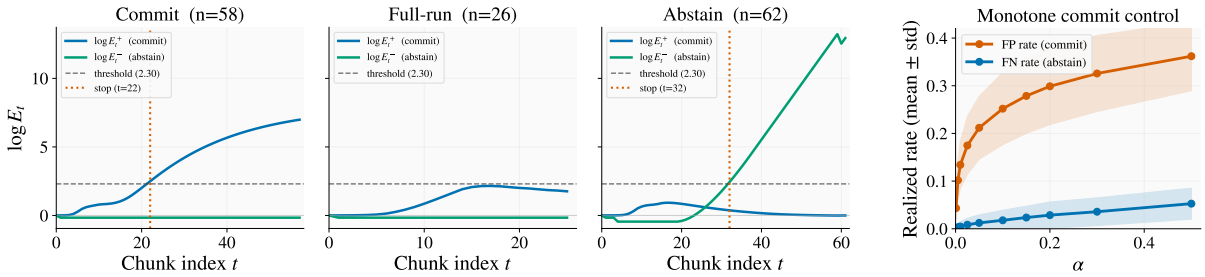
accuracy–token-use Pareto envelopes aggregated over all cells, for both upper-only stopping and dual stopping settings. Each curve sweeps over the method's parameter grid. Alongside our main recipe (stability signal), we also plug the signals evaluated in CT (EAT (Wang et al., 2026a) and confidence (Yang et al., 2026)) directly into our e-process framework, and conversely plug the stability signal into CT's framework, allowing a controlled signal–framework comparison.

As shown in Figure 1 and Table 1, the proposed method with stability signal clearly dominates CT across the majority of the token budget range in both settings, despite using zero calibration data where CT uses 50 validation trajectories per split. The advantage is consistent rather than concentrated at a single operating point. Using the signals evaluated in CT inside our e-process underperforms relative to the stability signal. This gap arises because EAT and confidence signals tend to plateau near the upper end of $[0, 1]$ once a reasoning chain is on track, producing a conditional mean that is approximately stationary under H_0^+. Such stationarity is well-suited to CT's sigmoid threshold, which compares the signal against a fixed level at each step. In our e-process, the wealth increment at each step measures how much the current signal deviates from its recent exponential moving average. When the signal is stationary, these deviations are near zero, the wealth accumulates negligibly, and the e-process rarely accumulates enough wealth to reach its crossing threshold. The stability signal avoids this problem by construction: it tracks changes in prediction consistency across chunks, so it continues to emit informative increments even late in the chain when EAT and confidence have already flattened. The truncation baseline, which stops generation at a fixed fraction of the token budget without any signal, is consistently dominated by the proposed method across the practical operating range, confirming that the accuracy gains stem from evidence-driven stopping rather than naive budget truncation. Per-combination Pareto curves are provided in Figure A in Appendix D.

4.3 Cross-Combination Deployment

We evaluate cross-cell generalization by fixing Qwen3-8B/AIME as the calibration source and deploying to all other model–dataset combinations. As shown in Figure 2, CT requires fitting a threshold τ on a held-out calibration set drawn from the deployment distribution, so reusing a source-calibrated τ on a new target combination breaks the calibration bound. The transferred τ is systematically miscalibrated: in the low-risk regime calibration is often infeasible under $n_{\text{val}} = 50$, and at higher risks the threshold is over-conservative, consuming more tokens than necessary.

The proposed method requires no calibration data at any stage: the commit and abstain decisions depend only on the current trajectory's e-process wealth, so the same α applies identically to any target cell without re-calibration. As a result, the proposed method achieves a consistently better token–accuracy trade-off across the full operating range in both cross-model and cross-dataset settings.



(a) Trajectory outcomes (α=0.1). (b) FP/FN rates vs. α.

Figure 3: E-process dynamics. **(a)** Three canonical trajectory outcomes at α=0.1: $\log E_t^+$ rises and crosses the threshold for a correct early prediction (*commit*); neither stream crosses for a hard instance that exhausts the full token budget (*full-run*); $\log E_t^-$ crosses first, correctly flagging a trajectory that would have produced a wrong terminal answer (*abstain*). **(b)** Realized commit FP rate and abstain FN rate as α is swept across all 12 model–dataset combinations; both decrease monotonically, confirming that (α_+, α_-) provide calibration-free monotone knobs on the two error types.

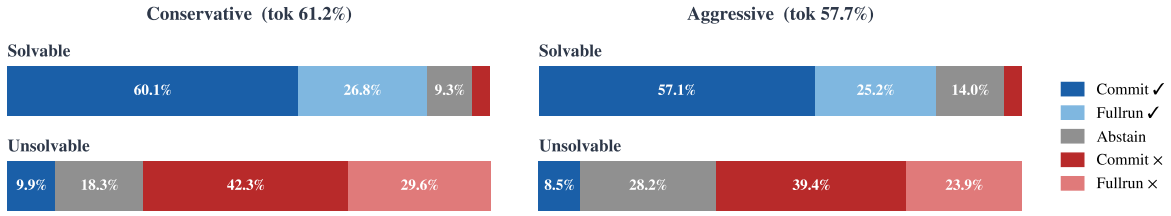


Figure 4: Stopping outcome distribution for Qwen3-8B on MATH-500 under two operating points. Each bar shows the fraction of trajectories in each outcome category, further decomposed into correct and incorrect predictions, broken down by problem solvability.

4.4 E-process Dynamics

Trajectory outcomes. We visualize representative trajectories to illustrate how the two wealth processes evolve across the three possible stopping outcomes. As shown in Figure 3a, we present three cases at α=0.1. In the commit case, $\log E_t^+$ rises steadily as the intermediate answers stabilize and crosses the threshold well before the final chunk, yielding an early correct prediction. In the full-run case, neither wealth process crosses the threshold: the model uses all available tokens and returns the terminal answer. In the abstain case, $\log E_t^-$ crosses the threshold while $\log E_t^+$ remains low throughout, flagging a trajectory on which the stability signal fails to converge.

Monotone commit control. Figure 3b plots the realized FP rate (commit stream) and FN rate (abstain stream) as α is swept across all 12 model–dataset combinations. Both rates decrease monotonically as α decreases, confirming that α_+ and α_- provide independently monotone knobs on the two error types without requiring calibration data.

Stopping outcome distribution. We analyze the distribution of trajectory outcomes at two operating points to characterize how the commit and abstain streams behave in practice. As shown in Figure 4, we report results for Qwen3-8B on MATH-500 under conservative and aggressive parameter settings, decomposing each outcome into whether it results in a correct or incorrect prediction. Early commit dominates in both settings, and the large majority of commits on solvable problems are correct, confirming that compute is reclaimed without sacrificing accuracy. On unsolvable problems, early commit captures a non-trivial fraction of correct predictions that would otherwise be lost to overthinking under a full-run policy, suggesting that the upper e-process fires precisely in windows where the model has already reached a correct intermediate answer before the reasoning chain deteriorates. The abstain stream exhibits a desirable asymmetry: unsolvable problems abstain at a substantially higher rate than solvable ones, indicating that the lower e-process selectively identifies likely-incorrect trajectories rather than firing indiscriminately. Moving from the conservative to the aggressive operating point

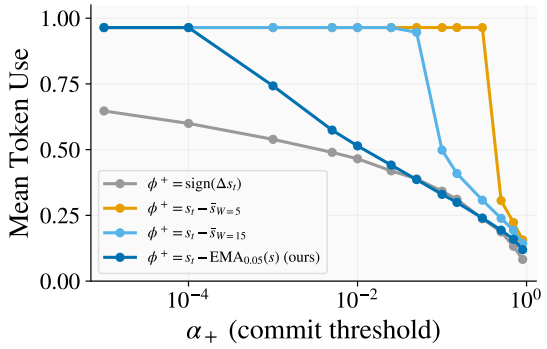


Figure 5: Four choices of the upper evidence increment ϕ^+ are compared by measuring mean token use as a function of α_+ (log scale), with the lower stream fixed to $\phi^- = \mathrm{sign}(-\Delta s_t)$ and results aggregated over all 12 model–dataset combinations.

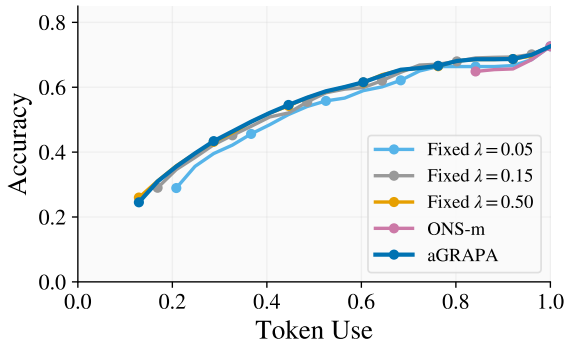


Figure 6: Five λ update strategies are compared on the token use–accuracy Pareto curve, with ϕ^+ fixed to $\mathrm{EMA}_{0.05}$ and results aggregated over all 12 model–dataset combinations.

raises the unsolvable-abstain rate while reducing mean token use, at the cost of a modest increase in false abstentions on solvable problems. Full-run cases — where neither e-process reaches its threshold within the token budget — account for a minority of trajectories in both settings.

4.5 Ablation Study

ϕ^+ construction. We ablate the choice of upper ϕ^+ by evaluating four constructions with the stability signal held fixed throughout. As shown in Figure 5, we report how mean token use responds to α_+ for each choice. Sign ($\phi^+ = \mathrm{sign}(\Delta s_t)$) responds to every α_+ value but saturates in the high-budget regime: the binary increment fires too readily, leaving no operating points there. Window-mean baselines ($\bar{s}_{W=5}$, $\bar{s}_{W=15}$) exhibit near-zero α_+ sensitivity across most of the practical range because the short window adapts quickly to the signal level, keeping $s_t - \bar{s}_W$ small and preventing wealth accumulation. $\mathrm{EMA}_{0.05}$ decreases monotonically across the full α_+ range, providing fine-grained budget control at every point.

λ update rule. We ablate the choice of betting fraction λ by comparing five strategies with ϕ^+ fixed to $\mathrm{EMA}_{0.05}$: three fixed values ($\lambda \in \{0.05, 0.15, 0.50\}$), ONS-m, and aGRAPA. As shown in Figure 6, fixed λ performance varies by value: λ=0.05 and λ=0.15 under-commit and fall below the adaptive frontier, while λ=0.50 performs comparably to aGRAPA but requires grid-searching over the λ space. ONS-m (Cutkosky and Orabona, 2018; Waudby-Smith and Ramdas, 2024), the Online Newton Step algorithm optimized for long observation horizons, collapses its λ to zero within the first few chunks of each reasoning trace, yielding operating points only near full token use and providing no early-stopping benefit; this shows that algorithms calibrated for long sequences degenerate when the observation horizon is as short as a reasoning trace. aGRAPA sustains non-zero λ throughout each trace and achieves a Pareto area matching the best fixed λ (λ=0.50) without requiring any hyperparameter search over the λ grid, making it our default choice.

5 Conclusion

We propose CES, a calibration-free framework for adaptive chain-of-thought stopping. CES accumulates sequential evidence through two parallel e-processes on a shared signal decomposed into position and signed velocity; the commit and abstain streams are anti-correlated by construction, so evidence for one decision is structurally evidence against the other. Under the null that stopping is unsafe, Ville's inequality gives an anytime-valid false-alarm bound controlled by a single pair (α_+, α_-). The bound holds at every step and under arbitrary continuation policies. The construction removes the calibration set from the deployment loop. Calibration-based stoppers tie their guarantees to a specific (model, dataset, budget) configuration and must be re-tuned whenever any factor shifts; with unseen models or distribution drift, calibration data is unavailable or stale. CES carries no such dependence: the same α transfers across models, datasets, and token budgets without held-out data.

Future work. Future work can consider composing the signal-level e-process with a downstream conformal layer to obtain a task-level coverage

guarantee on top of CES; instance-wise setting of α via test-time learning; extending the framework to alternative forms of test-time scaling, e.g., parallel reasoning chains; and improving the underlying per-chunk signal itself, e.g., by adding an auxiliary objective during post-training so that the increments ϕ_t^m correlate more directly with answer correctness.

Limitations

The primary limitation of our approach is that the false-alarm guarantee is signal-level: under the null of no systematic drift in ϕ_t^m, Ville's inequality bounds the probability that E_t^m crosses $1/\alpha_m$, but not the probability that an early commit yields a wrong answer. The correspondence between α_m and the task-level FP rate is monotone (Section 4.4) but not numerically tight, with the looseness depending on the trajectory distribution and on the correlation between s_t and answer correctness. Applications targeting a strict task-level risk should therefore layer a downstream calibration step on top of our framework.

Acknowledgements

This work was partly supported by the InnoCORE program of the Ministry of Science and ICT (AI Meta-Scientist, N10260110) and Artificial intelligence industrial convergence cluster development project funded by the Ministry of Science and ICT(MSIT, Korea)&Gwangju Metropolitan City.

References

Pranjal Aggarwal, Aman Madaan, Yiming Yang, and Mausam. 2023. 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*, pages 12375–12396, Singapore. Association for Computational Linguistics.

Anastasios N. Angelopoulos and Stephen Bates. 2023. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V Le, Christopher Ré, and Azalia Mirhoseini. 2024. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*.

Ashok Cutkosky and Francesco Orabona. 2018. Black-box reductions for parameter-free online learning in banach spaces. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 1493–1529. PMLR.

Yichao Fu, Xuewei Wang, Hao Zhang, Yuandong Tian, and Jiawei Zhao. 2026. Deep think with confidence. In *The Fourteenth International Conference on Learning Representations*.

Alex Gu, Baptiste Rozière, Hugh Leather, Armando Solar-Lezama, Gabriel Synnaeve, and Sida I. Wang. 2024. Cruxeval: A benchmark for code reasoning, understanding and execution. *arXiv preprint arXiv:2401.03065*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. 2024. OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3828–3850, Bangkok, Thailand. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, page 611–626, New York, NY, USA. Association for Computing Machinery.

Yiwei Li, Peiwen Yuan, Shaoxiong Feng, Boyuan Pan, Xinglin Wang, Bin Sun, Heda Wang, and Kan Li. 2024. Escape sky-high cost: Early-stopping self-consistency for multi-step reasoning. In *International Conference on Learning Representations*, volume 2024, pages 14751–14768.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's verify step by step. In *International Conference on Learning Representations*, volume 2024, pages 39578–39601.

Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Y. Tang, Manan Roongta, Colin Cai, Jeffrey Luo, Li Erran Li, Raluca Ada Popa, and Ion Stoica. 2025. Deepscaler: Surpassing o1-preview with a 1.5b model by scaling rl. Notion Blog.

Jingwei Ni, Ekaterina Fadeeva, Tianyi Wu, Mubashara Akhtar, Jiaheng Zhang, Elliott Ash, Markus Leippold, Timothy Baldwin, See-Kiong Ng, Artem Shelmanov,

and Mrinmaya Sachan. 2026. Reasoning with confidence: Efficient verification of LLM reasoning steps via uncertainty heads.

Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. 2023. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Glenn Shafer. 2021. Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 184(2):407–431.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2025. Scaling llm test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations*, volume 2025, pages 10131–10165.

Jean Ville. 1939. *Étude critique de la notion de collectif*. Monographies des probabilités, fasc. 3. Gauthier-Villars, Paris. Thèse, Faculté des Sciences de Paris.

Xi Wang, James McInerney, Lequn Wang, and Nathan Kallus. 2026a. Entropy after </think> for reasoning model early exiting. *Preprint*, arXiv:2509.26522.

Xi Wang, Anushri Suresh, Alvin Zhang, Rishi More, William Jurayj, Mehrdad Farajtabar, Daniel Khashabi, and Eric Nalisnick. 2026b. Conformal thinking: Risk control for reasoning on a compute budget. In *Forty-third International Conference on Machine Learning*.

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

Ian Waudby-Smith and Aaditya Ramdas. 2024. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):1–27.

Yasin Abbasi Yadkori, Ilja Kuzborskij, David Stutz, András György, Adam Fisch, Arnaud Doucet, Iuliya Beloshapka, Wei-Hung Weng, Yao-Yuan Yang, Csaba Szepesvári, and 1 others. 2024. Mitigating llm hallucinations via conformal abstention. *arXiv preprint arXiv:2405.01563*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Chenxu Yang, Qingyi Si, Yongjie Duan, Zheliang Zhu, Chenyu Zhu, Qiaowei Li, Minghui Chen, Zheng Lin, and Weiping Wang. 2026. Dynamic early exit in reasoning models. In *International Conference on Learning Representations*, volume 2026, pages 88170–88210.

Jinghan Zhang, Xiting Wang, Fengran Mo, Yeyang Zhou, Wanfu Gao, and Kunpeng Liu. 2025. Entropy-based exploration conduction for multi-step reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3895–3906, Vienna, Austria. Association for Computational Linguistics.

A Signal Definitions

We evaluate three per-chunk scalar signals $s_t \in [0, 1]$, each smoothed after raw computation with an exponential moving average (EMA) at $\alpha = 0.3$:

$$\bar{s}_t = 0.3 \cdot s_t + 0.7 \cdot \bar{s}_{t-1}, \quad \bar{s}_0 = s_0.$$

Confidence. The confidence signal is the geometric mean of per-token generation probabilities over chunk r_t:

$$s_t = \exp\left(\frac{1}{|r_t|} \sum_{i \in r_t} \log p_\theta(x_i \mid x_{<i})\right) \in (0, 1],$$

where $|r_t|$ is the number of tokens in the chunk. High values indicate that the model generates each token with high probability; near-zero values indicate uncertain generation. This signal is adopted from Yang et al. (2026).

EAT (Emission Answer Token). At each chunk boundary the model is forced to emit an intermediate answer; EAT records the entropy of the model distribution over the *first* emitted answer token (Wang et al., 2026a). We invert and normalize to $[0, 1]$ so that higher values indicate higher confidence:

$$s_t = 1 - \min\left(1, \frac{H_t}{\log 10}\right), \quad H_t = -\sum_v p_\theta(v \mid x, r_{1:t}) \log p_\theta(v \mid x, r_{1:t}),$$

where the sum is over the vocabulary and $\log 10$ normalizes against a ten-class uniform baseline.

Stability. The stability signal measures the persistence of the intermediate answer across consecutive chunks. Let $f_t(x)$ denote the normalized intermediate answer at chunk t. Define the run length as the number of consecutive chunks ending at t whose normalized answers all agree:

$$\text{run}_t = \max\{k \geq 1 : f_{t-k+1}(x) = \cdots = f_t(x)\}.$$

The raw signal is

$$s_t = \frac{\min(\text{run}_t,\ 8)}{8} \in [0, 1],$$

capped at 8 to bound sensitivity to long uninterrupted runs. High stability indicates the model has converged on an answer; low stability indicates the answer is still changing.

B Deeper Background

B.1 Setup

Let $(\mathcal{F}_t)_{t \geq 0}$ be a filtration with $\mathcal{F}_0$ trivial. A process (X_t) is *predictable* if X_t is $\mathcal{F}_{t-1}$-measurable for all $t \geq 1$. We work under a null hypothesis H_0, possibly composite, which constrains the conditional mean of the increment process specified in Eq. (D). Expectation $\mathbb{E}[\cdot]$ and probability $\Pr(\cdot)$ are taken under H_0 throughout.

B.2 E-processes and test supermartingales

A nonnegative process $(E_t)_{t \geq 0}$ adapted to $(\mathcal{F}_t)$ with $E_0 = 1$ is an *e-process* for H_0 if

$$\mathbb{E}[E_\tau] \leq 1 \tag{A}$$

for every stopping time τ. A *test supermartingale* under H_0 is a nonnegative process (M_t) with $M_0 = 1$ and $\mathbb{E}[M_t \mid \mathcal{F}_{t-1}] \leq M_{t-1}$. Every test supermartingale is an e-process; the converse holds for the constructions used here.

B.3 Ville's inequality

For any e-process (E_t) and any $\alpha \in (0,1)$,

$$\Pr\left(\sup_{t\geq 0} E_t \geq 1/\alpha\right) \leq \alpha. \tag{B}$$

The bound holds uniformly over t. Consequently, the stopping rule

$$\tau_\alpha = \inf\{\, t \geq 0 : E_t \geq 1/\alpha \,\} \tag{C}$$

controls the type-I error at level α regardless of when or how often E_t is inspected. This *anytime-validity* distinguishes e-processes from p-values, which require a pre-specified sample size.

B.4 Wealth construction by betting

Let (ϕ_t) be an $\mathcal{F}_t$-adapted increment process with $|\phi_t| \leq 1$ and

$$\mathbb{E}[\phi_t \mid \mathcal{F}_{t-1}] \leq 0, \qquad t \geq 1. \tag{D}$$

For any predictable (λ_t) with $\lambda_t \geq 0$ and $1 + \lambda_t \phi_t > 0$ a.s., the wealth process

$$E_t = \prod_{s=1}^{t} \big(1 + \lambda_s\, \phi_s\big), \qquad E_0 = 1, \tag{E}$$

is a nonnegative test supermartingale, and hence an e-process, for any H_0 satisfying (D). The interpretation is a gambler who, at round t, stakes a fraction λ_t of current wealth on the event $\{\phi_t > 0\}$.

B.5 aGRAPA Betting Fraction

The bet λ_t governs the power of the test: small λ_t accumulates evidence slowly, large λ_t risks rapid wealth loss when $\phi_t < 0$. A standard adaptive choice with no tuning parameter is aGRAPA (Waudby-Smith and Ramdas, 2024)

$$\lambda_t = \mathrm{clip}\left(\frac{\sum_{s<t} \phi_s}{1 + \sum_{s<t} \phi_s^2},\ 0,\ \lambda_{\max}\right), \tag{F}$$

which is $\mathcal{F}_{t-1}$-measurable and approximates the Kelly-optimal fraction online. With $|\phi_t| \leq 1$ and $\lambda_{\max} \leq 1/2$, the nonnegativity constraint $1 + \lambda_t \phi_t > 0$ holds automatically.

B.6 Composite nulls

The construction requires only the one-sided constraint (D) on the conditional mean of ϕ_t, not a fully specified distribution. The resulting e-process is therefore valid for any H_0 under which (D) holds, which is essential when the null is non-parametric or distributional assumptions on the underlying signal are weak.

C Theoretical Properties

This section establishes the anytime-valid guarantees of the construction. Notation follows Section 3; $m \in \{+,-\}$ indexes the two streams.

Lemma 1 (Bounded increments and predictable bets)**.** *For each $t \geq 1$:* (i) $\phi_t^+ \in [-1,1]$ *and* $\phi_t^- \in \{-1,0,1\}$*;* (ii) λ_t^m *is $\mathcal{F}_{t-1}$-measurable with $\lambda_t^m \in [0, \lambda_{\max}]$ and $\lambda_{\max} = 1/2$;* (iii) $1 + \lambda_t^m \phi_t^m \geq 1/2$ *almost surely.*

Proof. (i) and *(ii)* follow from (7)–(9). For *(iii)*, $\lambda_t^m \phi_t^m \geq -\lambda_t^m \geq -1/2$. □

Definition 1 (Signal-level nulls)**.**

$$H_0^+ : \mathbb{E}[\phi_t^+ \mid \mathcal{F}_{t-1}] \leq 0, \qquad H_0^- : \mathbb{E}[\phi_t^- \mid \mathcal{F}_{t-1}] \leq 0, \qquad \forall t \geq 1.$$

H_0^+ asserts that s_t does not systematically exceed its slow baseline; H_0^- asserts that s_t does not systematically decrease.

Lemma 2 (Supermartingale property)**.** *Under H_0^m, $(E_t^m)_{t\geq 0}$ is a nonnegative $(\mathcal{F}_t)$-supermartingale with $E_0^m = 1$.*

Proof. Nonnegativity follows from Lemma 1 (iii) by induction. Since $\lambda_t^m \geq 0$ is $\mathcal{F}_{t-1}$-measurable,

$$\mathbb{E}[E_t^m \mid \mathcal{F}_{t-1}] = E_{t-1}^m\big(1 + \lambda_t^m\, \mathbb{E}[\phi_t^m \mid \mathcal{F}_{t-1}]\big) \leq E_{t-1}^m$$

under H_0^m. □

Theorem 1 (Marginal type-I under the strict null)**.** *For any $\alpha_m \in (0,1)$,*

$$\Pr_{H_0^m}\big(\tau_m \leq T\big) \leq \alpha_m. \tag{G}$$

Proof. By Lemma 2 and Ville's inequality, $\Pr_{H_0^m}(\sup_{t\leq T} E_t^m \geq 1/\alpha_m) \leq \alpha_m$, which equals $\Pr_{H_0^m}(\tau_m \leq T)$ by definition of τ_m. □

The strict null requires zero mean drift of ϕ_t^m. In practice e_t lags s_t, so a slowly drifting signal can violate the assumption. We therefore state a relaxed bound that allows E_t^m to be a submartingale.

Theorem 2 (Marginal type-I under positive drift)**.** *Suppose $(E_t^m)_{t=0}^T$ is a nonnegative submartingale with $\mathbb{E}[E_T^m] \leq c_m$ for some $c_m \geq 1$. Then for any $\alpha_m \in (0,1)$,*

$$\Pr\big(\tau_m \leq T\big) \leq \alpha_m\, c_m. \tag{H}$$

Proof. Doob's maximal inequality gives, for any $a > 0$, $\Pr(\sup_{t\leq T} E_t^m \geq a) \leq \mathbb{E}[E_T^m]/a \leq c_m/a$. Set $a = 1/\alpha_m$. □

The constants c_+, c_- are reported empirically in Table F; α_m remains monotone in the false-decision probability.

Proposition 1 (Structural anti-correlation)**.** *For every $t \geq 1$,*

$$\mathrm{Cov}\big(\lambda_t^+\phi_t^+,\ \lambda_t^-\phi_t^- \,\big|\, \mathcal{F}_{t-1}\big) \leq 0. \tag{I}$$

Proof. Conditional on $\mathcal{F}_{t-1}$, the quantities $s_{t-1}, e_{t-1}, \lambda_t^+, \lambda_t^-$ are constants with $\lambda_t^m \geq 0$. Since $e_t = \gamma s_t + (1-\gamma)e_{t-1}$, we have $s_t - e_t = (1-\gamma)(s_t - e_{t-1})$, which is non-decreasing in s_t; hence $\phi_t^+ = \mathrm{clip}(s_t - e_t, -1, 1)$ is non-decreasing in s_t. By (8), $\phi_t^- = \mathrm{sign}(s_{t-1} - s_t)$ is non-increasing in s_t. Multiplying by the nonnegative constants λ_t^+, λ_t^- preserves monotonicity. The conditional covariance of two oppositely monotone functions of a single random variable is nonpositive (Harris's inequality, univariate case). □

Theorem 3 (Joint type-I via union)**.** *Under $H_0 := H_0^+ \cap H_0^-$,*

$$\Pr_{H_0}\big(\{\tau_+ \leq T\} \cup \{\tau_- \leq T\}\big) \leq \alpha_+ + \alpha_-. \tag{J}$$

Under the relaxed null of Theorem 2, the bound becomes $\alpha_+ c_+ + \alpha_- c_-$.

Proof. The false early-decision event is $\{\tau_+ \leq T\} \cup \{\tau_- \leq T\}$. Apply the union bound to (G) or (H). □

By Proposition 1, the per-step increments composing E_t^+ and E_t^- are conditionally anti-correlated, so simultaneous crossings are suppressed and (J) is typically loose in practice.

Scope of the guarantee. Theorems 1–3 bound the wealth-crossing probability under signal-level nulls, i.e., the absence of systematic drift in ϕ_t^m. This is not the same as the *task-level* event "committing now would yield a wrong answer." The task-level false-positive rate additionally depends on (i) the looseness of the bound for the trajectory distribution at hand and (ii) the correlation between s_t and answer correctness. Section 4.4 reports task-level rates separately: α_m is monotone in the task-level FP rate but not numerically equal to it, as is standard for any anytime-valid sequential test.

D Additional Experiment Results

D.1 Full Experiment Results

Figure A shows the token-use–accuracy Pareto envelopes broken down by all 12 model–dataset combinations (3 models $\times$ 4 datasets), for both the upper-only and dual stopping settings. The aggregate advantage reported in Figure 1 holds consistently at the per-cell level: the proposed method with the stability signal dominates CT-Conf and CT-EAT across the majority of the token budget range in every combination, despite using zero calibration trajectories where CT uses 50 per split.

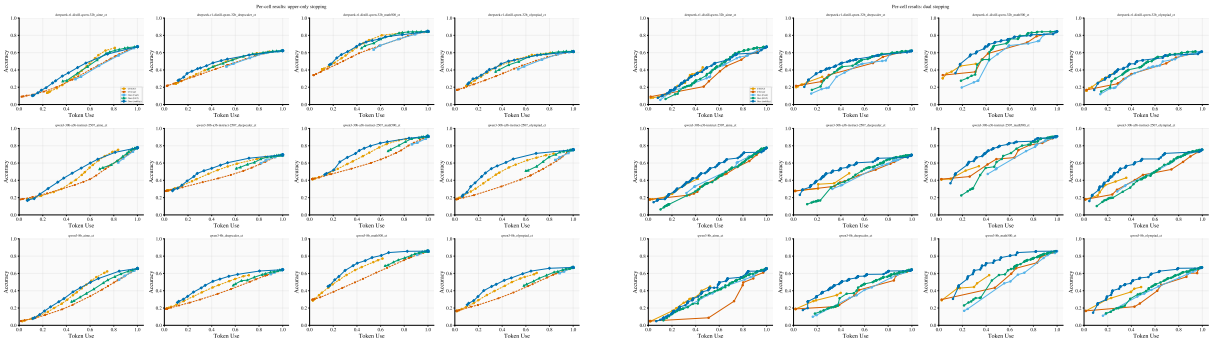


Figure A: Per-combination Pareto envelopes (upper-only, left; dual, right). Best viewed zoomed in.

Figure B shows the aggregated token-use–accuracy Pareto curve under cross-cell deployment in the dual stopping setting, averaged over all $12 \times 11 = 132$ source$\rightarrow$target pairs. CT-EAT transferred from a source cell to a different target cell degrades substantially relative to the same-cell reference: the calibration bound breaks when the deployment distribution differs from the calibration distribution. The proposed method requires no calibration data at any stage and uniformly dominates the cross-cell CT baseline across the full operating range.

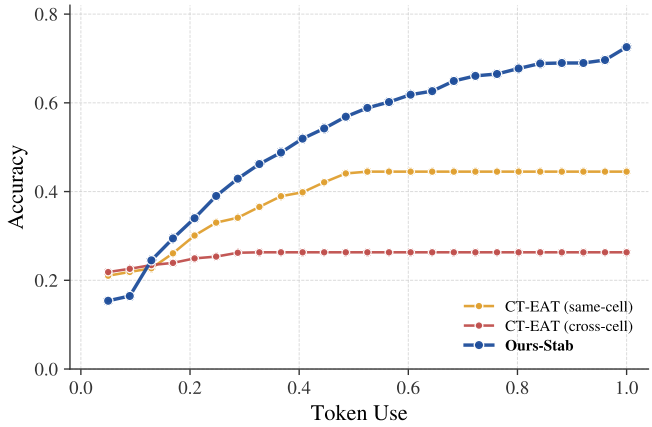


Figure B: Token-use vs. accuracy Pareto curves under cross-cell deployment (dual stopping), averaged over all $12 \times 11 = 132$ source$\rightarrow$target pairs. CT-EAT same-cell uses each target cell's own calibration data as a reference; CT-EAT cross-cell transfers the source-calibrated threshold to a different target cell. The proposed method (Ours-Stab) requires no calibration data and operates identically across all target cells.

D.2 Science Reasoning: GPQA-Diamond

We evaluate GPQA-Diamond (Rein et al., 2024), a graduate-level multiple-choice science benchmark containing 198 problems, on all three models from the main paper and two additional models: DeepSeek-R1-Distill-Llama-8B and DeepSeek-R1-Distill-Llama-70B (Guo et al., 2025).

Table A reports the dual-stopping accuracy envelope under normalized token-use constraints, aggregated over all five models. A dash indicates that no operating point exists at or below the corresponding budget.

Method	≤ 0.2	≤ 0.4	≤ 0.6	≤ 0.8
Truncation	0.385	0.447	0.503	0.568
CT-Conf	0.399	0.439	0.490	0.554
CT-EAT	0.407	0.456	0.489	0.506
CT-Stab	0.423	0.465	0.473	0.473
CES-Conf	0.112	0.209	0.375	0.511
CES-EAT	–	0.367	0.538	**0.616**
CES-Stab	**0.424**	**0.528**	**0.568**	0.594

Table A: Dual-stopping accuracy on GPQA-Diamond, aggregated over five models.

CES achieves the best accuracy at every reported budget. CES-Stab is strongest through token use ≤ 0.6, whereas CES-EAT becomes the best policy at ≤ 0.8.

D.3 Code Reasoning

We further evaluate CES on CRUXEval-O (Gu et al., 2024), a code-output prediction benchmark containing 800 problems. Each problem asks the model to reason about the value returned by a Python function. Because the target is a short Python literal, the intermediate-answer extraction and stability machinery used in the main paper applies without modification. We evaluate Qwen3-8B and DeepSeek-R1-Distill-Llama-8B and report the aggregate over the two model–dataset cells.

Method	≤ 0.2	≤ 0.4	≤ 0.6	≤ 0.8
Truncation	0.292	0.368	0.450	0.486
CT-Conf	0.286	0.348	0.425	0.469
CT-EAT	0.284	0.314	0.317	0.317
CT-Stab	**0.339**	0.339	0.339	0.339
CES-Conf	–	0.084	0.283	0.381
CES-EAT	–	0.251	0.424	0.488
CES-Stab	0.289	**0.447**	**0.484**	**0.492**

Table B: Dual-stopping accuracy on CRUXEval-O, aggregated over two models.

CT-Stab retains an advantage at the extreme ≤ 0.2 budget, whereas CES-Stab is the best policy from the practical token-use regime above approximately 0.3. Together with the other benchmarks, our evaluation therefore covers mathematical reasoning, science multiple choice, and code-output reasoning.

D.4 Other Backbones Results

We next evaluate two DeepSeek-R1-Distill (Guo et al., 2025) models using Llama-3.1 backbones. Each model is evaluated on the four mathematical benchmarks from the main paper together with GPQA-Diamond.

Method	≤ 0.2	≤ 0.4	≤ 0.6	≤ 0.8
Truncation	0.141	0.261	0.390	0.485
CT-Conf	0.203	0.298	0.400	0.480
CT-EAT	0.266	0.334	0.347	0.347
CT-Stab	0.261	0.340	0.364	0.364
CES-Conf	0.142	0.262	0.390	0.466
CES-EAT	0.173	0.346	0.474	**0.538**
CES-Stab	**0.276**	**0.416**	**0.485**	0.530

Table C: Dual-stopping accuracy for DeepSeek-R1-Distill-Llama-8B, aggregated over five datasets. Full-run accuracy across cells ranges from 0.475 to 0.744.

DeepSeek-R1-Distill-Llama-8B. CES-Stab is strongest through token use ≤ 0.6, while CES-EAT is best at ≤ 0.8.

Method	≤ 0.2	≤ 0.4	≤ 0.6	≤ 0.8
Truncation	0.236	0.350	0.503	0.620
CT-Conf	0.297	0.403	0.510	0.597
CT-EAT	0.355	0.438	0.447	0.452
CT-Stab	0.363	0.437	0.442	0.442
CES-Conf	–	0.274	0.478	0.598
CES-EAT	0.206	0.450	0.571	0.651
CES-Stab	**0.370**	**0.526**	**0.609**	**0.654**

Table D: Dual-stopping accuracy for DeepSeek-R1-Distill-Llama-70B, aggregated over five datasets. Full-run accuracy across cells ranges from 0.604 to 0.828; AIME full-run accuracy is 0.700.

DeepSeek-R1-Distill-Llama-70B. CES-Stab achieves the highest accuracy at every reported budget on the 70B model. Together with the main-paper models, these results extend the evaluation to five models spanning 8B–70B parameters and two backbone families.

E Computational cost analyses

E.1 Position-Aware FLOPs Accounting

Token count can overstate the marginal cost of intermediate probes because the prefix KV cache is reused. We therefore additionally report an analytic position-aware FLOPs estimate. For a token processed at context length L, we use

$$f(L) = 2N_{\text{active}} + 4n_{\text{layers}}n_{\text{heads}}d_{\text{head}}L, \tag{K}$$

where the first term accounts for weight-matrix multiplications and the second for KV-cache attention. For mixture-of-experts models, N_{active} denotes the number of active parameters. Since the reasoning prefix is already cached, the marginal probe FLOPs include only the forcing suffix and the at-most-24 greedily decoded answer tokens.

Across all cells, the normalized FLOPs overhead of probing is 0.43–0.83 relative to the original trace cost, compared with a normalized token overhead of 0.45–0.85. Token accounting is therefore slightly conservative relative to the analytic FLOPs accounting.

E.2 FLOPs-Normalized Pareto Results

Method	≤ 0.3	≤ 0.5	≤ 0.7
Truncation	0.277	0.435	0.548
CT-Conf (probe-free)	0.311	0.434	0.547
CT-EAT	0.323	0.390	0.430
CT-Stab	0.343	0.390	0.408
CES-Stab	**0.353**	**0.478**	**0.566**

Table E: Dual-stopping accuracy at normalized FLOPs budgets, aggregated over the 12 main-paper model–dataset cells.

CES-Stab remains the strongest policy at every reported FLOPs budget in Table E. The corresponding probe-inclusive token-axis results differ by approximately 0.01–0.02 accuracy over this range. Beyond approximately 0.8 of the full-trace budget, the setting is no longer a meaningful early-stopping regime: probe-free methods eventually overtake probe-based methods because probing overhead dominates when little generation is removed.

F Accuracy-Relevant Early-Commit Errors

For evaluating accuracy loss due to early stopping, we define an *accuracy-relevant commit false positive* as a trajectory satisfying all three conditions:

1. the commit stream crosses before the terminal boundary;

2. the intermediate prediction at the crossing is incorrect; and

3. the full-budget terminal prediction would have been correct.

Never-correct trajectories therefore do not contribute to this error count. Using all trajectories in each cell as the denominator, the accuracy-relevant FP rate is

$$0.26 \pm 0.08 \quad \text{at } \alpha_+ = 0.1, \tag{L}$$
$$0.13 \pm 0.05 \quad \text{at } \alpha_+ = 0.01. \tag{M}$$

When the denominator is restricted to final-correct trajectories, the corresponding rates are

$$0.363 \pm 0.127 \quad \text{at } \alpha_+ = 0.1, \tag{N}$$
$$0.192 \pm 0.078 \quad \text{at } \alpha_+ = 0.01. \tag{O}$$

The two normalizations are consistent because final-correct trajectories constitute 72.6% of all trajectories on average. Table F reports the complete per-cell mapping under both denominator conventions together with the empirical c_+ and c_- constants.

Model	Dataset	Final Acc.	$\alpha_+ = 0.1$		$\alpha_+ = 0.01$		c_+	c_-
			FP / All	FP / Final-Correct	FP / All	FP / Final-Correct		
Qwen3-8B	AIME	0.659	0.354	0.538	0.165	0.251	246	6.90
Qwen3-8B	MATH-500	0.858	0.168	0.196	0.068	0.079	451	0.80
Qwen3-8B	OlympiadBench	0.669	0.230	0.344	0.096	0.144	569	0.76
Qwen3-8B	DeepScaler	0.643	0.212	0.329	0.105	0.163	622	2.03
Qwen3-30B-A3B	AIME	0.777	0.361	0.464	0.188	0.242	419	119.90
Qwen3-30B-A3B	MATH-500	0.910	0.170	0.187	0.082	0.090	158	0.80
Qwen3-30B-A3B	OlympiadBench	0.754	0.309	0.409	0.171	0.226	840	0.79
Qwen3-30B-A3B	DeepScaler	0.692	0.242	0.350	0.135	0.196	494	105.58
R1-Distill-Qwen-32B	AIME	0.671	0.408	0.608	0.250	0.373	673	0.77
R1-Distill-Qwen-32B	MATH-500	0.846	0.160	0.189	0.094	0.111	481	0.82
R1-Distill-Qwen-32B	OlympiadBench	0.613	0.234	0.383	0.126	0.206	1118	0.80
R1-Distill-Qwen-32B	DeepScaler	0.622	0.227	0.365	0.135	0.218	912	0.82
Mean $\pm$ std		0.726	0.256 ± 0.079	0.363 ± 0.127	0.135 ± 0.050	0.192 ± 0.078		

Table F: Per-cell accuracy-relevant commit errors under two denominator conventions, together with empirical relaxed-null constants c_+ and c_-. "FP / All" divides the accuracy-relevant early-commit count by all trajectories in the cell, whereas "FP / Final-Correct" conditions the denominator on trajectories whose full-budget terminal answer is correct.

G Statement on the Use of Large Language Model

During the preparation of this work, the authors used a large language model for language editing assistance, including grammar, spelling, and phrasing. All content was reviewed and edited by the authors, who take full responsibility for the manuscript.

H License of Existing Assets

The models used in this work are released under permissive open-source licenses: Qwen3-8B and Qwen3-30B-A3B-Instruct-2507 are released under the Apache 2.0 license, and DeepSeek-R1-Distill-Qwen-32B is released under the MIT license. The benchmark datasets are used under their respective distribution terms: MATH-500 is derived from the MATH benchmark and the PRM800K MATH split, whose public repositories are MIT-licensed; OlympiadBench is distributed on Hugging Face under Apache 2.0 (with the official evaluation repository released under MIT); and DeepScaleR is released under the MIT license. For AIME, we use publicly available historical competition problems for academic evaluation only. All models and benchmark assets are used solely for research evaluation and in a manner consistent with their stated licenses and intended academic use.