When to Stop Thinking: Answer-Movement Certificates and Conditional Regret Guarantees

Anonymous authors
Paper under double-blind review

Abstract

Reasoning language models can improve their answers by thinking longer, but additional reasoning consumes resources and can also move an answer in the wrong direction. Efficient inference therefore requires deciding when to stop without access to the correct answer. We develop a framework that connects this decision to changes in the distribution of answers the model would give if stopped now. We derive answer-movement certificates that bound changes in any bounded answer score without correctness labels, yielding conditional regret bounds under cumulative movement budgets and sharper bounds near an optimal stopping policy. The analysis accommodates both beneficial and harmful reasoning and establishes why entropy stabilization alone is insufficient. We complement the theory with a calibrated entropy controller evaluated on 463 questions across four mathematics benchmarks using Qwen3.5-9B and Qwen3.5-27B. Across all eight model–benchmark comparisons, the controller reduces mean thought-plus-answer output tokens by 25–62% relative to horizon-limited native termination. Measured accuracy losses on competition tasks and the consistently lower cost of calibrated fixed budgets expose the limits of the entropy proxy. The framework provides explicit conditions for reliable stopping and a practical methodology for testing whether inexpensive stopping signals deliver useful accuracy–cost tradeoffs.

1 Introduction

Generating intermediate reasoning tokens gives a language model additional computation before it commits to an answer. Chain-of-thought prompting and reasoning-oriented training have made this an effective strategy for difficult problems (Wei et al., 2022; Kojima et al., 2022; DeepSeek-AI et al., 2025). Test-time scaling studies further show that the allocation of computation matters, rather than token count alone (Snell et al., 2024; Muennighoff et al., 2025). A fixed reasoning budget nevertheless treats an easy question and an unresolved hard question alike. An adaptive system must decide whether to continue using information available before it sees the correct answer.

The difficulty is that *movement*, *confidence*, and *improvement* are different properties. A model may confidently change from a correct answer to an incorrect one. It may temporarily stop changing before a later useful computation. A practical stopping signal should therefore be evaluated against a clearly specified object: how much the distribution of the model's eventual answer changes when another reasoning step is taken. We call the distribution obtained by interrupting thought and invoking a fixed answering procedure the *stop-now answer law*.

Our contribution is a framework connecting that law to utility and stopping cost. It does not require monotonic improvement. First, we decompose answer movement across complete reasoning traces into variation between traces and systematic change from direct answering, and distinguish movement from utility alignment. Second, a local KL certificate controls the magnitude of one-step utility change without correctness labels. We derive a global budget-dependent stopping bound that charges harmful late computation as well as missed beneficial computation, and a sharper bound under explicit proximity and intervening-movement assumptions. Third, a parity subproblem of permutation composition illustrates computational benefit under a carefully delimited architecture comparison. Finally, we specify and implement an evaluation that separates the exact certificate from

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.

tractable token-level proxies, fixes calibration before evaluation, and measures the cost of probing alongside output length.

These results identify both an opportunity and its conditions. Answer movement can constrain the value of computation for every bounded utility on a common answer space. Small recent movement by itself cannot certify that future computation is worthless. The cumulative-budget and proximity assumptions state precisely what must be added to obtain the corresponding stopping guarantees.

2 Related work

Reasoning and computation allocation. Chain-of-thought prompting (Wei et al., 2022; Kojima et al., 2022), reasoning-oriented models (DeepSeek-AI et al., 2025), and test-time compute allocation (Snell et al., 2024) motivate adapting the amount of computation to an input. The s1 procedure uses budget forcing to control generation length (Muennighoff et al., 2025). Our native-termination baseline simply permits the model to end its thought and then applies the common answer decoder; it does not implement that forcing procedure. Work on unfaithful explanations also cautions against treating a plausible rationale as direct evidence of the mechanism that produced a correct answer (Turpin et al., 2023).

Confidence-based early exit. Certaindex studies inference-compute allocation using uncertainty and reasoning progress (Fu et al., 2024). Entropy After Think uses the entropy following an inserted thought-ending marker for early exit (Wang et al., 2025). Our contribution is not the use of entropy or answer stability as a signal. It is the distinction between a utility-valid movement certificate and an inexpensive proxy, together with stopping guarantees whose assumptions are stated on the former. In particular, an entropy change need not bound the movement of a distribution.

Expressivity and optimal stopping. Hard-attention transformer results connect architectural restrictions to circuit complexity (Hao et al., 2022); chain-of-thought expressivity results explain how additional autoregressive computation can change computational power (Merrill & Sabharwal, 2023). We use a restricted parity example to illustrate benefit without claiming a new matched-resource transformer separation. The stopping formulation uses the finite-horizon Snell recursion, which we define directly, and the calibration bound follows from Hoeffding's inequality (Hoeffding, 1963). Our emphasis is the link between answer-law movement, possible utility change, and the assumptions needed to turn a local observation into a stopping decision.

3 Answer laws, utility, and the benefit of thought

For a fixed input x, we compare answer distributions over the same finite set of possible answers, $\mathcal{A}$. This set could be a list of candidate answers or all text responses from a finite vocabulary up to a specified length. We can use a simpler representation of these answers, but only if it preserves the distinctions that matter for scoring: answers grouped together must receive the same utility for x. We call such a representation *utility-sufficient*. The first token of a free-form mathematical answer generally does not meet this requirement. For example, "The answer is 4" and "The answer is 5" begin with the same token, even though one may be correct and the other incorrect.

For a fixed input x, let p_0 be the distribution of answers produced without a preceding reasoning trace. Let $Z \sim \mu_x$ be a complete reasoning trace sampled from the model, and let p_Z be the answer distribution after a particular trace z. We use the same answering procedure in both cases. Averaging over possible reasoning traces gives $\bar{p} = \mathbb{E}_Z[p_Z]$.

Let $U(a, x)$ be the score assigned to answer a for input x. Assume that the highest and lowest possible scores differ by at most $U_{\max} > 0$. For convenience, we subtract the lowest score for each input, so that $0 \le U(a, x) \le U_{\max}$. This shift changes neither the benefit of reasoning nor comparisons between stopping rules, and it keeps expected scores finite when averaging over inputs. The expected score under an answer distribution p is $U(p) = \sum_a p(a)U(a, x)$. We therefore define the benefit of reasoning as

$$\Delta(x) = U(\bar{p}) - U(p_0) = \mathbb{E}_Z[U(p_Z) - U(p_0)]. \tag{1}$$

Thus, reasoning helps when it increases the expected answer score and hurts when it decreases it. For exact-answer accuracy, correct answers score 1 and incorrect answers score 0, so $U_{\max} = 1$. These

scores define the benefit we evaluate; the controller does not need to know them while deciding when to stop.

Proposition 1 (Movement decomposition and utility ceiling)**.** *Assume* $J(x) = \mathbb{E}_Z \mathrm{KL}(p_Z \| p_0) < \infty$. *Under the joint law* $Z \sim \mu_x$, $Y \mid Z \sim p_Z$,

$$J(x) = I(Z; Y \mid x) + \mathrm{KL}(\bar{p} \| p_0), \tag{2}$$

$$|\Delta(x)| \le U_{\max} \, \mathrm{TV}(\bar{p}, p_0) \le U_{\max} \sqrt{\mathrm{KL}(\bar{p} \| p_0)/2} \le U_{\max} \sqrt{J(x)/2}. \tag{3}$$

KL uses natural logarithms and $\mathrm{TV}(p, q) = \frac{1}{2} \sum_a |p(a) - q(a)|$.

The first term in Equation (2) measures how much answers depend on the sampled thought; the second measures movement of the average answer law away from direct answering. Large variation between traces can coexist with zero average benefit. The ceiling limits the magnitude of benefit or harm; it does not predict its sign. If absolute continuity fails, the corresponding KL ceiling can be infinite, while the total-variation inequality remains valid.

A complementary lower bound isolates alignment with utility. Set $d = \mathbb{E}_Z \, \mathrm{TV}(p_Z, p_0)$, $g(Z) = U(p_Z) - U(p_0)$,

$$G_+ = \mathbb{E}[g(Z)_+], \qquad G_- = \mathbb{E}[(-g(Z))_+],$$

$$\rho = \mathbb{P}(g(Z) < 0), \qquad \kappa = \begin{cases} G_+/(U_{\max} d), & d > 0, \\ 0, & d = 0. \end{cases}$$

Here $v_+ = \max(v, 0)$, so both G_+ and G_- are nonnegative, and $\Delta = G_+ - G_-$. Utility transport implies $0 \le \kappa \le 1$.

Proposition 2 (Alignment floor)**.** *For every such utility and family of answer laws,*

$$\Delta(x) \ge \kappa U_{\max} d - U_{\max} \rho. \tag{4}$$

This bound can be informative when answer movement is aligned with better answers and harmful traces are rare. Its parameters involve utility: it is an explanatory decomposition, not a label-free input gate. A nonpositive lower bound does not imply that thought is harmful. In particular, uncertainty or confidence alone cannot identify ρ, which concerns a decline in conditional expected utility rather than the event that one sampled answer is wrong.

4 A RESTRICTED COMPUTATIONAL EXAMPLE

Consider S_5 composition, $f_n(\pi_1, \ldots, \pi_n) = \pi_1 \circ \cdots \circ \pi_n$. On the subfamily with independent uniform $\pi_i \in \{e, \sigma\}$, $\sigma = (1\,2)$, the answer encodes the parity of the n input bits.

Proposition 3 (Restricted lower bound and sequential algorithm)**.** *Consider a constant-precision, unique-hard-attention transformer family of fixed depth and width whose answer computation, including* $O(1)$ *autoregressive thought passes, is representable by constant-depth, polynomial-size Boolean circuits. Its deterministic accuracy on the uniform parity subfamily is at most* $1/2 + o(1)$. *The same conclusion holds for randomized predictors if every fixed-randomness realization obeys common depth and polynomial-size bounds of this circuit class. Separately, a sequential finite-state algorithm computes* f_n *exactly in* n *updates using 120 states.*

The circuit restriction is the relevant hypothesis, supported for the specified hard-attention setting by architectural expressivity results (Hao et al., 2022). It does not include unrestricted soft attention, growing precision, or an external sampling primitive that computes outside this class. The lower bound follows from the average-case parity limitation of constant-depth polynomial-size circuits (Håstad, 1987). The algorithm stores the running group product and updates it using the fixed Cayley table. A transformer implementation would additionally need to address the next input and preceding state. The proposition therefore claims neither a matched-width transformer construction nor an $\Omega(n)$ lower bound on necessary thought tokens.

The example makes Equation (4) concrete. A direct head uniform on $\{e, \sigma\}$ has accuracy $1/2$, while the sequential answer is correct with probability one. Thus $d = 1/2$, $\kappa = 1$, $\rho = 0$, and the floor equals the benefit $1/2$. If both laws are mixed with the uniform law on S_5 with weight $\varepsilon \in (0, 1)$, they have full support and benefit $(1 - \varepsilon)/2$. At thought-token price $\lambda = c/n$ with $0 < c < 1/2$, the deterministic sequential algorithm has net utility $1 - c > 1/2$. This is a specified algorithmic comparison, not a claim about the experimental language models.

5 LOCAL MOVEMENT AND OPTIMAL STOPPING

Let $\mathcal{F}_k$ contain the input and the first k thought tokens. At each prefix, apply a fixed stop-now answering procedure, obtaining an $\mathcal{F}_k$-measurable answer law q_k on $\mathcal{A}$. Let

$$V_k = \sum_a q_k(a)U(a,x), \qquad W_k = V_k - \lambda k, \qquad 0 \le k \le H, \tag{5}$$

where $H \ge 1$ is an integer horizon and $\lambda > 0$ prices a thought token. Changing the answering format changes the intervention and hence the process. The oracle and proposed rule must use the same process, utility and price.

Why this is an intervention process. V_k is the expected value of answering at prefix k, not a declaration that $V_k = \mathbb{E}[R \mid \mathcal{F}_k]$ for a single fixed terminal reward R. The latter would be a bounded martingale, and optional stopping would give no expected reward gain from delaying, while charging a positive cost. Computation is useful here because a later prefix can change the answering procedure's output distribution.

Proposition 4 (Local movement certificate)**.** *For $C_k = \mathrm{KL}(q_k \| q_{k-1})$, whenever this KL is finite,*

$$|V_k - V_{k-1}| \le U_{\max}\,\mathrm{TV}(q_k, q_{k-1}) \le U_{\max}\sqrt{C_k/2}. \tag{6}$$

More generally, any observable, nonnegative A_k satisfying $|V_k - V_{k-1}| \le U_{\max}\sqrt{A_k/2}$ is a valid local certificate for the given utility process.

The exact KL is independent of correctness labels if the full answer laws can be evaluated. Evaluating them can nevertheless be intractable for free-form answers. The certificate bounds the step just taken; a small value is neither a correctness claim nor a bound on the value of all future steps. The choice $A_k = 2\,\mathrm{TV}(q_k, q_{k-1})^2$ also satisfies local validity and is at most two, without a support assumption, but still requires the full answer laws.

Assumption 1 (Uniformly bounded surprisal)**.** There is a deterministic $b < \infty$ such that $q_k(a) \ge e^{-b}$ for every answer $a \in \mathcal{A}$, every $0 \le k \le H$, and every reachable prefix, almost surely.

This sufficient condition gives $C_k \le b$ and $\sum_{k=1}^{H} C_k \le Hb$. For a finite candidate head with logits in $[-R, R]$ and softmax temperature $T > 0$, one may take $b = \log|\mathcal{A}| + 2R/T$. This example concerns candidate-answer logits; it does not assign the same bound to probabilities of complete autoregressive strings. Greedy decoding and nucleus truncation can create zero probabilities even when underlying logits are finite. In particular, the study's top-p sampler does not establish Assumption 1, and adding a numerical floor to a truncated diagnostic distribution cannot establish it for the true answer law.

Oracle. For stopping times taking values in $\{0, \dots, H\}$ and adapted to $\mathcal{F}_k$, define

$$S_H = W_H, \qquad S_k = \max\{W_k, \mathbb{E}[S_{k+1} \mid \mathcal{F}_k]\}, \qquad \tau^\star = \inf\{k : S_k = W_k\}.$$

The finite-horizon recursion gives $\mathbb{E}W_{\tau^\star} = \sup_\tau \mathbb{E}W_\tau$. This oracle knows the conditional law and utility function but does not see future random tokens. A hindsight best prefix on one realized trajectory is a different object.

Assumption 2 (Local validity and cumulative budget)**.** The $\mathcal{F}_k$-measurable certificates $A_k \ge 0$ obey

$$|V_k - V_{k-1}| \le U_{\max}\sqrt{A_k/2}, \qquad \sum_{k=1}^{H} A_k \le B$$

almost surely, for a deterministic finite B.

Define the ideal first-low-certificate rule

$$c = 2\lambda^2/U_{\max}^2, \qquad \widehat{\tau} = \min\big(\inf\{1 \le k \le H : A_k \le c\}, H\big), \tag{7}$$

with $\inf \varnothing = +\infty$.

Theorem 5 (Budget-dependent stopping guarantee)**.** *Under Assumption 2,*

$$0 \le \mathbb{E}[W_{\tau^\star} - W_{\widehat{\tau}}] \le \min\left\{U_{\max} + \lambda H, \frac{U_{\max}^2 B}{\lambda} + 2\lambda\right\}. \tag{8}$$

The proof separately charges missed improvement on early-stop paths and both utility decline and token cost on late-stop paths. No monotonicity or martingale condition on V_k is imposed. The budget controls how much adverse movement can occur after a low certificate; it need not imply a small numerical bound. For example, the generic choice $B = Hb$ may be too large to improve on the trivial finite-horizon cap. Small recent movement with unrestricted later movement yields no strong guarantee from this argument.

A conditional local guarantee. For an integer $1 \le w \le H$, let

$$\widehat{\tau}^{(w)} = \min\left(\inf\left\{w \le k \le H : \frac{1}{w}\sum_{i=k-w+1}^{k} A_i \le c\right\}, H\right).$$

Assumption 3 (Proximity and intervening movement)**.** Almost surely, $|\widehat{\tau}^{(w)} - \tau^\star| \le w$, and, writing $s = \min(\widehat{\tau}^{(w)}, \tau^\star)$, $t = \max(\widehat{\tau}^{(w)}, \tau^\star)$, $\sum_{i=s+1}^{t} A_i \le wc$. Local validity holds as in Assumption 2.

Theorem 6 (Conditional window guarantee)**.** *Under Assumption 3,* $\mathbb{E}[W_{\tau^\star} - W_{\widehat{\tau}^{(w)}}] \le 2w\lambda$.

No global budget is needed for this theorem. The window test itself does not establish proximity to the oracle or control certificate mass on an unobserved future segment. These are separate substantive premises, plausible only where useful computation and the stopping signal are locally aligned. Delayed insight, a long quiet interval followed by a decisive step, or utility oscillation can violate them. If the complete pathwise premises of either theorem hold only on an event of probability at least $1 - \delta$, its expected bound can be increased by $\delta(U_{\max} + \lambda H)$, using the unconditional payoff range. A pooled disagreement rate across observed steps is not such a trace-level failure probability.

6 From the certificate to a practical controller

A practical system cannot enumerate all answer strings at every thought token. Our implementation inserts the same thought-closing and answering cue used by the final answer decoder and requests the top-K next-token log probabilities. Let r_k denote the distribution reconstructed from this support after the chosen filtering and normalization. The primary signal is the signed entropy drop

$$D_k = \mathrm{H}(r_{k-1}) - \mathrm{H}(r_k). \tag{9}$$

Raw entropy, absolute entropy drop, reconstructed KL and reconstructed total variation define alternative controls; the empirical evaluation focuses on the primary rule. These token distributions are diagnostic objects, not the full q_k of Theorem 4. The probe requests untruncated log probabilities with greedy sampling settings, whereas actual answers use temperature 0.6 and top-$p = 0.95$; a matched prompt does not make these laws identical.

Entropy is insufficient. For $r_{k-1} = (0.99, 0.01)$ and $r_k = (0.01, 0.99)$, $D_k = 0$ but $\mathrm{KL}(r_k \| r_{k-1}) = 0.98 \log 99 \approx 4.50$. If the first answer is correct, accuracy drops by 0.98. Absolute values or clipping negative drops cannot repair this failure. In general, with $\widetilde{q}_{k-1} = \mathbb{E}[q_k \mid \mathcal{F}_{k-1}]$,

$$\mathrm{H}(q_{k-1}) - \mathbb{E}[\mathrm{H}(q_k) \mid \mathcal{F}_{k-1}] = \mathrm{H}(q_{k-1}) - \mathrm{H}(\widetilde{q}_{k-1}) + \mathbb{E}[\mathrm{KL}(q_k \| \widetilde{q}_{k-1}) \mid \mathcal{F}_{k-1}]. \tag{10}$$

Only with predictive-mixture consistency, $q_{k-1} = \widetilde{q}_{k-1}$, does the conditional expected entropy drop reduce to the information term. That same consistency makes the bounded linear-utility process a martingale, so it cannot simply be imposed as an explanation for expected computational improvement.

What an approximation guarantee would require. If approximations r_{k-1}, r_k live on the same utility-sufficient answer space and $\mathrm{TV}(q_j, r_j) \le \varepsilon_j$, the triangle inequality gives

$$|V_k - V_{k-1}| \le U_{\max}\left[\sqrt{\mathrm{KL}(r_k \| r_{k-1})/2} + \varepsilon_k + \varepsilon_{k-1}\right]. \tag{11}$$

This identifies the missing ingredient for a certified approximation. Dropping token mass, renormalizing a top-K support, or representing a full answer by its first token does not supply these error bounds. A floor added to the union of observed supports makes reconstructed KL finite and reproducible; it changes the diagnostic law and does not bound omitted full-answer probability.

Operational policy. The primary controller probes at the empty prefix and then every 32 emitted thought tokens, uses filtered top-20 support, and stops when two consecutive eligible observations have $D_k \leq \gamma$. Eligibility starts after 64 thought tokens, so the earliest threshold-triggered stop is at token 96. Missing or empty filtered support does not trigger a stop. Native termination and the horizon also stop generation. The final answer is decoded at the actual stopping prefix, using the common answering intervention. Generated EOS tokens are counted when token IDs identify them; metadata-only finish notifications add no output token. Minimum length, patience, sparse probing, and native termination are explicit operational choices. They are absent from the ideal rule in Equation (7), so its theorem is not automatically a guarantee for this controller.

Calibration and its guarantee. Thresholds and fixed caps are selected on separate calibration inputs. The implemented objective is accuracy minus η times *thought plus answer* output tokens, with the output-token price η distinct from the theory's thought-token price λ. Probe overhead is measured separately and excluded from this calibration objective. If answer length is policy-dependent, the resulting policy comparison need not coincide with maximizing Equation (5). For a policy π, define its payoff on one episode and its expected payoff by

$$R(\pi) = U(A_\pi, X) - \eta(T_\pi + N_\pi), \qquad \overline{R}(\pi) = \mathbb{E}[R(\pi)],$$

where X is the input, A_π is the generated answer, and T_π and N_π count thought and answer output tokens, respectively. The expectation in $\overline{R}(\pi)$ averages over a fresh input and generation randomness.

Proposition 7 (Finite-grid calibration)**.** *Let Π be a fixed family of M policies, and let $R_i(\pi)$ be the payoff measured for policy π on calibration episode i. Suppose the n calibration episodes are i.i.d. and independent of evaluation, with $\mathbb{E}[R_i(\pi)] = \overline{R}(\pi)$ for every $\pi \in \Pi$. Assume all measured payoffs lie in a common interval of width L. Define*

$$\widehat{R}_n(\pi) = \frac{1}{n}\sum_{i=1}^{n} R_i(\pi), \qquad \widehat{\pi} \in \arg\max_{\pi\in\Pi} \widehat{R}_n(\pi), \qquad \pi^*_{\mathrm{grid}} \in \arg\max_{\pi\in\Pi} \overline{R}(\pi).$$

For $0 < \delta < 1$, with probability at least $1 - \delta$ over the calibration episodes,

$$\overline{R}(\pi^*_{\mathrm{grid}}) - \overline{R}(\widehat{\pi}) \leq 2L\sqrt{\frac{\log(2M/\delta)}{2n}}.$$

Consequently, for any $\epsilon > 0$, $n \geq 2L^2\epsilon^{-2}\log(2M/\delta)$ suffices for a gap in expected payoff of at most ϵ with the same probability.

For the implemented output-token objective one may use $L = U_{\max} + \eta(H + A_{\max})$, with the declared answer cap $A_{\max}$. The comparison is to the best policy in the fixed grid, not the unrestricted stopping oracle. Calibration sample sizes are design choices, not a claim that they satisfy a particular target (ϵ, δ). A finite benchmark split also does not by itself prove an i.i.d. population model.

Calibration replays each candidate on a sampled native thought trajectory and independently decodes answers at its exact stop. This yields the candidate's on-policy expectation only if continuing thought transitions are invariant to the remaining cap until truncation, and side probes do not change the thought state or transition law. Shared calibration trajectories may correlate policies' estimates; the union bound permits that dependence. Independent episodes across inputs and unbiased per-policy measurements remain necessary. Live evaluation executes each selected rule directly, so it does not rely on replay to produce its test outcomes.

7 Experiments

We evaluate the accuracy and output cost of the calibrated entropy controller against native termination and a calibrated fixed budget. The evaluation covers two model sizes and four mathematics benchmarks, with identical questions and scoring rules across methods within each benchmark.

7.1 Models and datasets

We use hosted Qwen3.5-9B and Qwen3.5-27B models, abbreviated as 9B and 27B. The serving interface provides top-20 token log probabilities and a 262,144-token context limit. Model identifiers, tokenizer revisions and serving metadata are recorded in Section B.

Table 1: Evaluation design per model. All three methods use the same 463 evaluation questions. Thought and answer limits apply separately; AIME and HMMT use MATH calibration.

Dataset	Calib.	Eval.	Thought cap	Answer cap
GSM8K	8 train	200 test	256	512
MATH-500	16	200	4,096	4,096
AIME 2026	0	30	8,192	8,192
HMMT Feb. 2026	0	33	8,192	8,192

GSM8K provides grade-school arithmetic questions (Cobbe et al., 2021); MATH-500 covers more varied mathematical reasoning (Hendrycks et al., 2021; Lightman et al., 2023). We sample eight GSM8K calibration questions from the training split and 200 evaluation questions from the test split. MATH-500 is divided into disjoint pools of 100 calibration and 400 evaluation questions, from which we sample 16 and 200 questions, respectively. MATH evaluation is stratified by source difficulty level and subject, covering all five levels and seven subjects. GSM8K evaluation includes equal numbers from question word-length quartiles, a measure of input breadth rather than a validated difficulty scale. The 30 AIME and 33 HMMT questions comprise all available items in the specified competition snapshots. Calibration and evaluation inputs have no exact normalized-question overlap; the sampling details and scope of inference appear in Section B.

7.2 Baselines, calibration, and generation

The *native* reference permits natural thought termination up to the benchmark horizon. The *fixed* reference selects a thought cap using the same calibration objective as the primary controller. Its candidate caps are $\{0, 64, 256\}$ for GSM8K and $\{0, 256, 1024, 4096\}$ for MATH, with zero denoting direct answering. Calibration selects 64 thought tokens for GSM8K and zero for MATH for both models. MATH-selected thresholds and fixed caps transfer unchanged to AIME and HMMT. A zero thought cap still permits the common answer decoder to produce an explanation.

The primary controller follows the policy in Section 6. We select its threshold from $\{0.001, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5\}$ by maximizing calibration accuracy minus output cost at $\eta = 10^{-4}$. Threshold ties use a canonical policy identifier; fixed-cap ties favor the shorter cap. All selections use calibration inputs and are held fixed during evaluation. Table 7 reports the selected primary policies.

Each method produces one answer per evaluation question. Generation uses temperature 0.6, top-$p = 0.95$, and the thought and answer limits in Table 1. Methods share the same answering cue and answer cap within each benchmark. The native and primary thought trajectories are bounded by the stated horizon; native termination can occur earlier. Truncated answers remain in the evaluation denominator.

Thought generation is served in chunks of at most 256 tokens, with exact token IDs preserved across continuations. Logical output cost counts thought tokens consumed by the stopping rule plus generated answer tokens. Server-reported usage also includes any returned chunk tokens that the controller does not consume. We measure these unused tokens, probe calls, request counts and elapsed time separately. Latency includes queueing, network delay and probing, with up to eight concurrent episode workers per model. These measurements do not identify GPU computation or unobserved server run-ahead.

7.3 Scoring and statistical analysis

We use benchmark-specific automatic graders consistently across methods. Mathematical answer extraction normalizes explicit final values and single-variable assignments without consulting the gold answer. Unparsed, unsupported and timed-out comparisons are scored incorrect and reported separately; the reported accuracies do not include manual adjudication.

The question is the unit of statistical analysis. Paired 95% confidence intervals use 10,000 bootstrap resamples of shared question IDs, including for ratios of mean output cost. Exact paired accuracy tests use discordant outcomes, with Bonferroni correction across the 16 primary model–dataset–

reference comparisons. We assess noninferiority separately at a two-percentage-point margin using the conservative bound in Section B.1. The evaluation is exploratory because the cohort also informed protocol development, as detailed in Section B.

7.4 Accuracy and output cost

Table 2 compares the primary controller with each reference on GSM8K and MATH. Accuracy is the primary controller's percentage correct; differences are primary minus reference in percentage points. Savings measure the percentage reduction in mean thought-plus-answer output tokens relative to the reference. Brackets give paired 95% confidence intervals.

Table 2: Accuracy and output cost on GSM8K and MATH.

Model	Data	Ref.	n	Acc.%	Δ pp [95% CI]	Saving% [95% CI]
27B	GSM8K	Native	200	89.5	0.5 [-3.5, 4.5]	25.2 [23.3, 27.1]
27B	GSM8K	Fixed	200	89.5	-2.5 [-6.0, 1.0]	-19.8 [-23.4, -16.4]
27B	MATH	Native	200	63.0	-2.0 [-9.0, 5.0]	62.4 [58.7, 66.1]
27B	MATH	Fixed	200	63.0	-13.0 [-19.5, -6.5]	-41.4 [-53.0, -30.7]
9B	GSM8K	Native	200	86.0	1.0 [-4.0, 6.0]	30.3 [28.5, 32.2]
9B	GSM8K	Fixed	200	86.0	1.5 [-3.5, 6.5]	-9.3 [-12.7, -5.9]
9B	MATH	Native	200	57.5	-4.0 [-11.0, 3.0]	61.4 [58.0, 64.8]
9B	MATH	Fixed	200	57.5	-18.5 [-25.5, -12.0]	-44.5 [-57.6, -32.2]

On GSM8K, the primary controller reduces mean output cost relative to native termination for both models, while the paired accuracy intervals include zero. On MATH, it also reduces output cost relative to native termination, but the calibrated fixed reference achieves higher accuracy at lower cost. Both MATH accuracy deficits relative to the fixed reference remain significant after correction for multiple comparisons (Table 6). Thus, savings relative to native termination alone do not demonstrate an advantage over calibrated fixed budgets.

7.5 Transfer to competition mathematics

Section C.1 reports transfer from MATH calibration to AIME and HMMT. Output cost is lower than native termination in all four model–dataset comparisons, but accuracy is also lower, with each pointwise paired accuracy interval excluding zero. The calibrated fixed reference uses fewer output tokens than the primary controller in every competition comparison. The small contest cohorts and substantial answer truncation limit the precision and interpretation of these estimates. Adjusted exact-test results are reported separately in Table 6.

7.6 Completion quality and serving cost

Table 5 reports thought-horizon hits, answer truncation and unresolved grading alongside accuracy for all three methods. The automatic grader leaves 32% of primary MATH answers unresolved for each model. Across the three methods and both models, 233 of 378 competition answers reach the answer cap. These rates limit conclusions about mathematical capability despite the larger budgets assigned to difficult benchmarks.

The serving measurements in **??** show why output length and latency must be reported separately. On GSM8K, the primary controller uses fewer logical output tokens than native termination but has higher mean latency for both models. On MATH and the competition benchmarks, its mean latency is lower than native termination, while the calibrated fixed reference is faster still. These comparisons describe the recorded serving workload, including probe overhead and queueing.

8 Discussion and limitations

Answer-distribution movement connects computational benefit to stopping, but its usefulness depends on the answer representation and future movement constraints. Exact full-answer certificates may be intractable; truncated token probes do not close this gap. Bounded surprisal can supply a finite budget but need not hold under the experimental sampler. The total-variation alternative avoids this support premise while still requiring full-answer laws.

A cumulative budget controls worst-case continuation through a potentially loose bound. The sharper window guarantee requires proximity to the oracle; it does not prove that a practical entropy policy achieves it. Both permit utility decreases. Delayed useful computation, confidently wrong answers, format changes, support mismatch and distribution shift can defeat the proxy.

The empirical scope is limited by automatic grading, finite generation budgets, small calibration and competition cohorts, and evaluation of only two sizes in one model family. Development-related reuse of the evaluation cohort is described in Section B.

No primary comparison establishes noninferiority at the specified two-percentage-point margin. The observed output savings relative to native termination therefore do not establish preservation of answer quality. The stronger performance of calibrated fixed budgets on MATH and the competition accuracy losses emphasize the need to evaluate inexpensive stopping signals against both accuracy and serving cost. The conditional full-answer guarantees specify assumptions for reliable stopping; the token-level heuristic does not establish those assumptions.

AI use statement

OpenAI Codex assisted with editing the manuscript's structure and references, conversion to the ICLR format, implementing and checking the hosted-model adapter and experiment runner, and generating manuscript tables.

Reproducibility statement

Section A provides the proofs, and Sections B and 7 specify the datasets, disjoint calibration and evaluation partitions, policy grids, generation settings, scoring rules and estimators. The retained artifacts include input identifiers, raw outputs, token-level probe records, serving measurements, seeds and implementation fingerprints. All performance comparisons use the same complete evaluation cohort and automatic scoring metric. Replication of the hosted models additionally depends on access to the serving environment.

References

Karl Cobbe et al. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168*, 2021.

DeepSeek-AI et al. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint arXiv:2501.12948*, 2025.

Yichao Fu et al. Efficiently Scaling LLM Reasoning with Certaindex. *arXiv preprint arXiv:2412.20993*, 2024.

Yiding Hao, Dana Angluin, and Robert Frank. Formal Language Recognition by Hard Attention Transformers: Perspectives from Circuit Complexity. *arXiv preprint arXiv:2204.06618*, 2022.

Johan Håstad. *Computational Limitations of Small-Depth Circuits*. MIT Press, 1987.

Dan Hendrycks et al. Measuring Mathematical Problem Solving With the MATH Dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Wassily Hoeffding. Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Takeshi Kojima et al. Large Language Models are Zero-Shot Reasoners. *arXiv preprint arXiv:2205.11916*, 2022.

Hunter Lightman et al. Let's Verify Step by Step. *arXiv preprint arXiv:2305.20050*, 2023.

William Merrill and Ashish Sabharwal. The Expressive Power of Transformers with Chain of Thought. *arXiv preprint arXiv:2310.07923*, 2023.

Niklas Muennighoff et al. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025.

Charlie Snell et al. Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters. *arXiv preprint arXiv:2408.03314*, 2024.

Miles Turpin et al. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. *arXiv preprint arXiv:2305.04388*, 2023.

Xi Wang, James McInerney, Lequn Wang, and Nathan Kallus. Entropy After `</Think>` for reasoning model early exiting. *arXiv preprint arXiv:2509.26522*, 2025.

Jason Wei et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *arXiv preprint arXiv:2201.11903*, 2022.

A PROOFS

A.1 UTILITY TRANSPORT AND MOVEMENT DECOMPOSITION

For $0 \le U \le U_{\max}$, splitting the signed measure $p - q$ into its positive and negative parts gives $|U(p) - U(q)| \le U_{\max}\,\mathrm{TV}(p, q)$. Pinsker's inequality in natural logarithms gives $\mathrm{TV}(p, q) \le \sqrt{\mathrm{KL}(p\|q)/2}$; an infinite KL yields a vacuous upper bound. For finite J, expand

$$\mathbb{E}_Z \sum_a p_Z(a) \log \frac{p_Z(a)}{p_0(a)} = \mathbb{E}_Z \sum_a p_Z(a) \log \frac{p_Z(a)}{\bar{p}(a)} + \sum_a \bar{p}(a) \log \frac{\bar{p}(a)}{p_0(a)}.$$

The first term is $I(Z; Y \mid x)$ and the second is $\mathrm{KL}(\bar{p}\|p_0)$. Nonnegativity and utility transport prove Theorem 1. For Theorem 2, $\Delta = G_+ - G_-$ and $G_- \le U_{\max}\mathbb{P}(g(Z) < 0) = U_{\max}\rho$. Also $G_+ \le \mathbb{E}|g(Z)| \le U_{\max} d$, proving the stated range of κ. When $d = 0$, the answer laws agree almost surely and the benefit and both parts vanish.

A.2 SUPPORT, LOCAL VALIDITY, AND APPROXIMATE LAWS

Under Assumption 1,

$$\mathrm{KL}(q_k\|q_{k-1}) = \sum_a q_k(a) \log q_k(a) - \sum_a q_k(a) \log q_{k-1}(a) \le b,$$

since the first sum is nonpositive and $-\log q_{k-1}(a) \le b$. For the candidate-head softmax, each numerator is at least $e^{-R/T}$ and the denominator is at most $|\mathcal{A}|e^{R/T}$, giving the asserted b. Utility transport proves Theorem 4. Applying the total-variation triangle inequality through r_k, r_{k-1} and then Pinsker to their KL proves Equation (11). The common answer space and utility-sufficiency premises are needed for the first utility-transport step.

A.3 SNELL RECURSION AND THE GLOBAL BOUND

Backward induction shows that S_k is the maximal conditional expected stopped payoff from time k: stopping yields W_k and continuing yields $\mathbb{E}[S_{k+1} \mid \mathcal{F}_k]$. The first equality time attains this maximum in the finite horizon. All quantities are integrable because $V_k \in [0, U_{\max}]$ and $k \le H$.

We prove a pathwise bound against any reference time τ, then take $\tau = \tau^\star$. If $\widehat{\tau} < \tau$, let $g = \tau - \widehat{\tau}$ and $B_g = \sum_{i=\widehat{\tau}+1}^{\tau} A_i$. Cauchy–Schwarz gives

$$W_\tau - W_{\widehat{\tau}} \le U_{\max}\sqrt{gB_g/2} - \lambda g \le \frac{U_{\max}^2 B_g}{8\lambda} \le \frac{U_{\max}^2 B}{8\lambda}.$$

The middle bound maximizes the preceding expression over real $g \ge 0$, so also covers integer gaps. If $\widehat{\tau} > \tau$, every intervening $i < \widehat{\tau}$ has $A_i > c$. For those steps,

$$|V_i - V_{i-1}| + \lambda \le U_{\max}\sqrt{A_i/2} + \lambda \le \frac{U_{\max}^2 A_i}{\lambda}.$$

Indeed, each of the two terms is at most $U_{\max}^2 A_i/(2\lambda)$ when $A_i \ge c$. A last step with $A_{\widehat{\tau}} \le c$ contributes at most 2λ. If the horizon instead forces a high-certificate final step, charge it to B in the same way. Thus the late-stop gap is at most $U_{\max}^2 B/\lambda + 2\lambda$. Equal times have zero gap. The unconditional payoff range is at most $U_{\max} + \lambda H$, establishing the second cap. Taking expectations and using optimality of $\tau^\star$ proves Theorem 5.

A.4 WINDOW AND EXCEPTIONAL-EVENT BOUNDS

Let $d = |\widehat{\tau}^{(w)} - \tau^\star| \le w$. By local validity and the assumed intervening mass,

$$|V_{\widehat{\tau}^{(w)}} - V_{\tau^\star}| \le U_{\max}\sqrt{dwc/2} \le w\lambda.$$

The absolute cost difference is at most $w\lambda$, proving Theorem 6. If an event E contains the full hypotheses of a pathwise bound G and $\mathbb{P}(E^c) \le \delta$, then $\mathbb{E}[W_{\tau^\star} - W_{\widehat{\tau}}] \le G + \delta(U_{\max} + \lambda H)$ by splitting the expectation over E and E^c. No conditioning that would change the oracle or its information structure is needed.

A.5 Entropy identity and calibration

Conditionally on $\mathcal{F}_{k-1}$, expanding KL gives

$$\mathbb{E}[\mathrm{KL}(q_k \| \widetilde{q}_{k-1}) \mid \mathcal{F}_{k-1}] = \mathrm{H}(\widetilde{q}_{k-1}) - \mathbb{E}[\mathrm{H}(q_k) \mid \mathcal{F}_{k-1}].$$

Adding $\mathrm{H}(q_{k-1}) - \mathrm{H}(\widetilde{q}_{k-1})$ proves Equation (10). Predictive-mixture consistency implies $\mathbb{E}[V_k \mid \mathcal{F}_{k-1}] = V_{k-1}$ for the fixed bounded utility, as stated.

For calibration, Hoeffding's inequality (Hoeffding, 1963) and a union bound over the M fixed policies give $|\widehat{R}_n(\pi) - \overline{R}(\pi)| \le a$ simultaneously for every $\pi \in \Pi$, where $a = L\sqrt{\log(2M/\delta)/(2n)}$, with probability at least $1 - \delta$. On that event,

$$\overline{R}(\pi^*_{\mathrm{grid}}) - \overline{R}(\widehat{\pi}) \le \overline{R}(\pi^*_{\mathrm{grid}}) - \widehat{R}_n(\pi^*_{\mathrm{grid}}) + \widehat{R}_n(\widehat{\pi}) - \overline{R}(\widehat{\pi}) \le 2a,$$

using the maximizing definition of $\widehat{\pi}$. Solving $2a \le \epsilon$ proves Theorem 7. Repeated answer samples within an episode reduce conditional sampling noise but do not multiply the number of independent calibration inputs.

A.6 Computational example details

On inputs $\pi_i \in \{e, \sigma\}$ the product is σ exactly when the number of transpositions is odd. A constant number of passes in the stated circuit-representable transformer family composes to constant depth and polynomial size. Testing whether its output equals σ yields a parity predictor whose accuracy is at least that of the original S_5-valued predictor: any output outside $\{e, \sigma\}$ was already wrong. The average-case parity bound for this circuit class therefore proves the accuracy limit. In the randomized case the same bound must hold uniformly over fixed-randomness realizations before averaging.

For the sequential algorithm initialize $P_0 = e$ and set $P_k = P_{k-1} \circ \pi_k$ via the finite multiplication table. Induction gives $P_k = \pi_1 \circ \cdots \circ \pi_k$. This uses 120 possible recurrent states and n input-indexed transitions. Addressing those transitions in a transformer is an additional construction, not supplied by the finite-state algorithm. The accuracy, total variation and floor calculations in Section 4 follow directly from its deterministic correct answer and the specified uniform direct head.

B Protocol details and reproducibility

Controller and calibration. The primary controller uses filtered top-20 support, a minimum of 64 thought tokens, probe cadence 32 and patience two. Missing or empty filtered support does not trigger a stop. Calibration samples one native thought trajectory per input and independently decodes an answer at each distinct prefix required by the threshold candidates and fixed-cap grid. Candidates share an answer measurement only when they select the same prefix under the same answering intervention. The MATH-selected fixed cap transfers unchanged to both competition datasets. Patience counts consecutive eligible probe observations; it does not establish a theorem-level bound on intervening full-answer movement.

Data and measurement records. The study contains 48 calibration trajectory records and 2,778 evaluation episodes: 463 questions, two models and three methods. Each evaluation record retains the generated answer, thought tokens, probe observations, grading status and request-level usage. Data, model settings, measurement code and analysis code have recorded fingerprints, and record checksums establish the inputs to each reported table. Failed requests retain separate records. All planned evaluation episodes are included. Unresolved grading cases and a fixed 1% sample of other answers are retained for inspection; no manual adjudication is included in the reported metric.

Scope of inference. The evaluation cohort was also used during protocol development to identify response truncation and answer-extraction failures. Token budgets and extraction rules were then fixed for all measurements reported here. Threshold and fixed-budget selection used disjoint calibration inputs, and calibration-only engineering checks are excluded from performance estimates. The evaluation is therefore exploratory: the confidence intervals and multiplicity corrections describe the specified comparisons but do not account for protocol choices informed by development on this cohort.

Complexity coverage. MATH evaluation includes 19, 37, 45, 47 and 52 questions at source levels 1–5. Subject counts are 49 algebra, 14 counting/probability, 16 geometry, 39 intermediate algebra, 23 number theory, 34 prealgebra and 25 precalculus. Quotas are proportional within nonempty level–subject strata, with at least one question per stratum. GSM8K contributes 50 questions per word-length quartile. Within each declared stratum, selection uses seeded input identifiers and the disjoint source partitions. No question is removed on the basis of measured correctness or runtime.

B.1 Statistical estimation and noninferiority

One generation is retained per evaluation input and method. Uncertainty resamples input clusters using 10,000 percentile-bootstrap draws at 95% confidence. Paired comparisons use shared input IDs and resample the paired clusters, including for ratios of mean output cost. These intervals describe input-resampling uncertainty; small competition sets and the stratified benchmark sample limit precision and population interpretation. These descriptive intervals are not a design-based estimate for an unspecified deployment population.

A formal noninferiority decision is kept separate from these descriptive intervals. The comparison family contains $2 \times 1 \times 4 \times 1 \times 2 = 16$ primary comparisons: model, format, dataset, price, and native/calibrated-fixed reference. For input-level accuracy differences in $[-1, 1]$, the conservative lower bound is

$$\mathrm{LCB} = \max\left\{-1, \bar{d} - \sqrt{\frac{2\log(16/0.05)}{n}}\right\}.$$

Noninferiority at the two-percentage-point margin requires $\mathrm{LCB} > -0.02$, under independent sampled-input assumptions. This correction is intentionally conservative and can be uninformative at the available sample sizes. A nonsignificant accuracy difference, overlapping intervals, or a favorable point estimate does not establish noninferiority.

Power and exact paired tests. With one binary correctness outcome per input, we test paired accuracy differences using the exact two-sided binomial test on discordant pairs. The 16 primary tests use Bonferroni-adjusted p-values; this tests an accuracy difference, not noninferiority. Wins and losses are reported in Section C.3. A hypothetical normal-approximation power calculation assumes independent pairs and a discordant fraction of 0.30, rather than estimating it from evaluation outcomes. At $n = 200$, 80% power requires an accuracy difference of approximately 10.9 percentage points at unadjusted $\alpha = 0.05$, or 14.7 points at $\alpha = 0.05/16$. These are scenario calculations, not a significance guarantee. The 30/33-question competition cohorts are less precise. Bootstrap intervals and exact tests use the question as the unit; models and token probes do not multiply the number of independent evaluation questions.

B.2 Assumptions and their empirical status

Table 3: Assumptions and empirical limitations.

Premise	When it can hold	Failure or limitation here
Common utility-sufficient answer space	Finite candidates or complete capped strings with utility defined on each outcome.	First-token equality need not imply equal correctness of completed answers.
Bounded utility	Accuracy and other normalized finite-range scores.	Unbounded losses need a different transport or integrability argument.
Full-answer local validity	Exact KL or TV of stop-now laws; same-space approximations with proved error bounds.	Filtered top-K entropy and reconstructed token KL supply no full-answer certificate by themselves.
Finite cumulative budget	Bounded surprisal across reachable prefixes, or a directly justified TV/movement budget.	Top-p creates zeros; observed reconstructed budget quantiles do not prove a pathwise bound.

Premise	When it can hold	Failure or limitation here
Window proximity and mass	A regime in which stopping is already close to the oracle and intervening movement is controlled.	Quiet periods followed by delayed insight, oscillations and proxy mismatch may violate both.
Unbiased replay calibration	Passive probes and identical causal thought transitions until the selected cap.	Budget-dependent generation behavior, stateful probe side effects or decoder mismatch can bias replay estimates.
Independent calibration/evaluation inputs	Separate population samples under a fixed protocol.	Disjoint benchmark splits prevent direct fitting leakage but do not establish population independence or absence of training contamination.
Measured resource equivalence	An explicitly instrumented cost model including the resources being claimed.	Logical output counts exclude probe computation and unobserved server run-ahead.

C Additional result tables

These tables complement the GSM8K and MATH comparisons in Section 7.4, reporting competition transfer, completion quality, paired accuracy tests, calibration selections and serving measurements for the same evaluation protocol.

C.1 Competition transfer

AIME and HMMT use the MATH-selected threshold and fixed cap at output-token price $\eta = 10^{-4}$. Each primary policy is compared with horizon-limited native termination and the calibrated fixed reference under the same answer-generation settings.

Table 4: Competition accuracy and output cost under transferred calibration.

Model	Data	Ref.	n	Acc.%	Δ pp [95% CI]	Saving% [95% CI]
27B	AIME	Native	30	16.7	-36.7 [-53.3, -20.0]	35.5 [29.2, 41.3]
27B	AIME	Fixed	30	16.7	-16.7 [-33.3, 0.0]	-52.2 [-84.1, -30.3]
27B	HMMT	Native	33	15.2	-30.3 [-48.5, -12.1]	40.5 [35.8, 44.6]
27B	HMMT	Fixed	33	15.2	-27.3 [-42.4, -12.1]	-22.3 [-38.5, -11.3]
9B	AIME	Native	30	10.0	-26.7 [-46.7, -10.0]	41.2 [36.0, 45.3]
9B	AIME	Fixed	30	10.0	-16.7 [-33.3, 3.3]	-59.6 [-98.1, -34.1]
9B	HMMT	Native	33	6.1	-18.2 [-33.3, -3.0]	47.3 [43.1, 52.3]
9B	HMMT	Fixed	33	6.1	-30.3 [-45.5, -15.2]	-20.3 [-39.9, -5.1]

C.2 Completion and grading quality

"Thought cap" is the fraction reaching the benchmark's maximum thought horizon, not the fraction reaching a shorter selected fixed cap. "Answer cap" uses the server's length-termination metadata. "Unresolved" includes extraction failures, symbolic timeouts and unsupported expressions, all scored incorrect under the fixed metric. A truncated answer may already contain a scorable final value; the fixed extraction rule still applies, and such outputs remain flagged. Accuracy uses every declared input, without filtering on completion or grading status.

Table 5: Generation limits and automatic grading by model, dataset and method.

Model	Data	Method	n	Acc.%	Thought cap%	Answer cap%	Unresolved%
27B	GSM8K	Primary	200	89.5	5.0	3.0	1.0
27B	GSM8K	Native	200	89.0	100.0	2.0	1.0
27B	GSM8K	Fixed	200	92.0	0.0	2.0	0.0
27B	MATH	Primary	200	63.0	0.0	21.5	32.0
27B	MATH	Native	200	65.0	63.0	8.0	30.0
27B	MATH	Fixed	200	76.0	0.0	5.0	18.5
27B	AIME	Primary	30	16.7	0.0	70.0	0.0
27B	AIME	Native	30	53.3	100.0	30.0	3.3
27B	AIME	Fixed	30	33.3	0.0	36.7	0.0
27B	HMMT	Primary	33	15.2	0.0	81.8	63.6
27B	HMMT	Native	33	45.5	100.0	42.4	45.5
27B	HMMT	Fixed	33	42.4	0.0	63.6	36.4
9B	GSM8K	Primary	200	86.0	1.0	2.0	0.0
9B	GSM8K	Native	200	85.0	100.0	3.0	0.5
9B	GSM8K	Fixed	200	84.5	0.0	3.5	0.5
9B	MATH	Primary	200	57.5	0.0	28.5	32.0
9B	MATH	Native	200	61.5	64.5	19.0	30.0
9B	MATH	Fixed	200	76.0	0.0	8.0	16.5
9B	AIME	Primary	30	10.0	0.0	93.3	0.0
9B	AIME	Native	30	36.7	100.0	60.0	0.0
9B	AIME	Fixed	30	26.7	0.0	40.0	3.3
9B	HMMT	Primary	33	6.1	0.0	90.9	66.7
9B	HMMT	Native	33	24.2	100.0	69.7	45.5
9B	HMMT	Fixed	33	36.4	0.0	57.6	42.4

C.3 Exact paired accuracy comparisons

Wins/losses count inputs on which only the primary/reference answer is correct. Adjusted values cover all 16 primary model/dataset/reference comparisons; significance is assessed at 0.05 using unrounded values. A nonsignificant result does not establish equal accuracy or noninferiority.

Table 6: Exact paired accuracy tests with correction for multiple comparisons.

Model	Data	Format	Ref.	n	Wins/losses	Exact p	Adjusted p
27B	GSM8K	free	Native	200	9/8	1.0000	1.0000
27B	GSM8K	free	Fixed	200	5/10	0.3018	1.0000
27B	MATH	free	Native	200	23/27	0.6718	1.0000
27B	MATH	free	Fixed	200	11/37	0.0002	0.0036
27B	AIME	free	Native	30	0/11	0.0010	0.0156
27B	AIME	free	Fixed	30	1/6	0.1250	1.0000
27B	HMMT	free	Native	33	1/11	0.0063	0.1016
27B	HMMT	free	Fixed	33	0/9	0.0039	0.0625
9B	GSM8K	free	Native	200	15/13	0.8506	1.0000
9B	GSM8K	free	Fixed	200	15/12	0.7011	1.0000
9B	MATH	free	Native	200	20/28	0.3123	1.0000
9B	MATH	free	Fixed	200	9/46	$< 10^{-4}$	$< 10^{-4}$
9B	AIME	free	Native	30	1/9	0.0215	0.3438
9B	AIME	free	Fixed	30	2/7	0.1797	1.0000
9B	HMMT	free	Native	33	1/7	0.0703	1.0000
9B	HMMT	free	Fixed	33	0/10	0.0020	0.0312

C.4 Calibration

The primary thresholds below were selected on the calibration partitions at $\eta = 10^{-4}$. The calibrated fixed caps were 64 thought tokens on GSM8K and zero on MATH for both models; MATH selections also apply to AIME and HMMT.

Table 7: Primary policies selected by calibration.

Model	Data	Format	Threshold	Patience	Cadence	Cal. utility
27B	GSM8K	free	0.200	2	32	0.968
27B	MATH	free	0.050	2	32	0.628
9B	GSM8K	free	0.500	2	32	0.972
9B	MATH	free	0.050	2	32	0.472