SEQUENTIAL ANYTIME VALID TESTS FOR EFFICIENT INFERENCE-TIME ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Inference-time alignment methods, such as Best-of-N (BoN), improve language model outputs by generating many candidates per prompt and selecting among them using a reward model. The cost is largely wasted, since a fixed budget is applied uniformly across prompts, even though the computation required by different prompts varies widely. Adaptive methods halt generation once the sampled answers appear settled, but their thresholds are tuned heuristics with no control over the probability of stopping at the wrong answer. Instead, we ask for a rule that takes that probability – a permitted error level – as input, and returns the number of tokens spent to meet it. To build such a rule, we aggregate candidates by their final answers using a reward-weighted self-consistency (RWSC) score and treat the decision of when to stop as a sequential test over the resulting clusters. Wald's sequential probability ratio test (SPRT) is the natural tool for such a test, and when applied to the RWSC clusters, it yields SPRT-RWSC, which performs well empirically. Yet its guarantee does not hold here, since the leading answer shifts as sampling proceeds, and the run is continuously monitored. Our main contribution replaces Wald's test with an anytime-valid construction that resolves this issue: it maintains a betting supermartingale (an e-process) for every answer the policy produces, bounding the probability of ever certifying a non-dominant answer by α, uniformly over time and across all answers, however many the policy produces. We call the resulting rule e-RWSC: the RWSC scores drive the bets, which affect only how quickly a dominant answer $a^\star$ is certified, not the validity of the guarantee; and certification takes $O\big(\log(1/\alpha)/d_{\mathrm{KL}}(\theta_{a^\star}||1/2)\big)$ samples. Experiments across several language models, datasets, and reward models confirm the predicted $1/d_{\mathrm{KL}}$ scaling of per-response cost, and show that accuracy improves over baselines at substantially lower token cost.

1 INTRODUCTION

Inference-time alignment has emerged as the practical alternative to fine-tuning, in which one repeatedly generates responses from the existing language model (e.g., a policy) and selects among them. Best-of-N (Beirami et al., 2025) selection scores each response with a reward model and returns the one with the highest score. Self-consistency (Wang et al., 2022) takes a majority vote over final answers. Both are effective, but expensive: their cost is N full generations per prompt, and N must be large enough for the hardest prompt in the workload. Since inference is the dominant cost of deployed language models, the token budget spent on selection should be minimized.

A fixed budget of N is uniform across prompts, while the computation required by prompts varies enormously. On a prompt where the policy immediately concentrates on a single answer, ten samples are enough, whereas on a prompt on which the policy remains genuinely divided, even a thousand samples are not sufficient. Fixed-budget selection, therefore, results in both over- and underspending of tokens. Adaptive methods address this by halting generation once the sampled answers look settled, e.g., stopping when a posterior over answer frequencies concentrates (Aggarwal et al., 2023; Wan et al., 2025), when the marginal value of another sample falls below its cost (Kalayci et al., 2025), or when a budget estimated from a short burn-in is exhausted (Raman et al., 2026). In every case, the halting rule is a tuned threshold with no control over the probability of stopping on the wrong answer. Moreover, they must be re-tuned for each new policy, reward model, and benchmark.

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.

One would instead like a stopping rule that takes an error level as input and returns the required token budget, e.g., a practitioner specifies how often they are willing to commit to a wrong answer, and the generation continues until the threshold is met. The rule must be defined based on a computable statistic of the generations. Since many distinct reasoning traces can yield the same final answer, we group responses by their final answers and frame the stopping rule at the answer level. We introduce Reward-Weighted Self-Consistency (RWSC), which scores each cluster by the softmax-weighted mass of its members' reward scores, so that a cluster is favored both for containing many responses and for highly rewarded ones. Let a^* denote the highest-scoring cluster (the leading answer) at the current step. Framing stopping as the question – *is $a^\star$ genuinely the dominant answer, or merely ahead by chance?* – makes it a sequential hypothesis testing problem. Wald's sequential probability ratio test (SPRT) is the classic solution: as each new response arrives, one accumulates evidence for or against the leader and stops once it crosses a boundary (Wald, 1945). It performs strongly in practice, but remains a heuristic here, without any rigorous error-control guarantee.

Wald's guarantee assumes hypotheses that are fixed in advance, tested by a single terminal decision, and none of these holds here. Writing θ_a for the probability that a fresh response yields answer a, the rates defining the test are estimated from the same responses the test then examines, rather than fixed beforehand. The leading answer a^* is itself a function of the sampled responses, so the parameter under test, θ_{a^*}, shifts whenever the leader does, forcing the test to restart. And the statistic is checked after every response, not at a single stopping point. Each of these voids the error control that Wald's boundaries nominally provide. We resolve this with anytime-valid inference (Ramdas et al., 2023), whose guarantees hold under continuous monitoring and arbitrary stopping. We maintain a separate betting supermartingale for each final answer the policy produces, instantiating one when that answer first appears and assigning it a weight from a fixed, summable sequence. Since the weights sum to 1, they split a single error budget across answers, regardless of how many arise. Ville's inequality (Ville, 1939) then bounds by α the probability that a non-dominant answer ($\theta_a \leq 1/2$) is ever certified, uniformly over time and over a growing set of answers. This addresses each failure of SPRT. A change in the leading answer is irrelevant, since every answer accumulates its own evidence. Continuous monitoring is an inherent property of Ville's inequality, so checking the boundary after each response incurs no additional cost.

Validity requires only that the bets be predictable, each computed from earlier responses before the next one is sampled. The reward model sets each bet based on the current RWSC score of the answer cluster, which buys efficiency when the scores are accurate. The gain has a precise form: under the log-optimal Kelly betting (Kelly, 1956), evidence for a dominant answer accrues at a rate $d_{\mathrm{KL}}(\theta_a||1/2)$, so certification takes about $\log(1/\alpha)/d_{\mathrm{KL}}(\theta_a||1/2)$ sampled responses. An accurate RWSC score bets near-optimally and approaches this rate, while an inaccurate or hacked score bets poorly and slows the certification. But it never crosses the boundary in expectation under the null, since validity does not depend on the estimate being correct.

The efficiency of the stopping rule, therefore, adapts to the prompt. The number of tokens spent scales with its difficulty: an answer the policy strongly favors ($\theta_a > 1/2$) is certified in a few responses, whereas one near the decision boundary $\theta_a = 1/2$ takes many, often more than the budget allows, in which case the procedure certifies nothing and falls back to standard aggregation. Either way, the practitioner fixes the error-control probability α once, and the number of responses adjusts itself to each prompt. We summarize our *contributions* as follows.

(i) We cast early stopping as a sequential hypothesis-testing problem and introduce SPRT-RWSC, which applies Wald's test to determine when the leading answer has converged. It operates on Reward-Weighted Self-Consistency (RWSC): an answer-level aggregation rule that scores answer clusters by reward-weighted mass, and recovers majority voting and Best-of-N as extreme cases. It works well in practice, but Wald's error guarantee does not hold in this setting (Section 2).

(ii) We propose E-RWSC, an anytime-valid stopping rule that accumulates evidence over all answers through separate e-processes. It controls the Type-I error (certifying an answer that is not dominant under the policy) at a user-specified level α, uniformly over time, and over an unknown set of answers. The guarantee holds for any predictable choice of bets, so the reward model, which sets them through RWSC scores, governs only efficiency and never validity. Under a plug-in estimate of the log-optimal Kelly betting, certifying the dominant answer (when one exists) takes $O\big(\log(1/\alpha)/d_{\mathrm{KL}}(\theta_{a^*}||1/2)\big)$ responses as $\alpha \to 0$. So, the number of tokens spent scales with the prompt's difficulty (Section 3).

(iii) An empirical study on 4 Language Models (Qwen2.5-3B, Qwen2.5-7B, Llama-3-8B, Phi-3-Mini-4K) with 4 Reward Models (ArmoRM, FsfairX, Skywork V1, Skywork V2) over 4 Benchmarks (GSM8K, MATH500, MMLU, CruxEval) showing that realized certification error stays below the nominal level, that per-response cost follows the predicted $1/d_{\mathrm{KL}}$ scaling with prompt difficulty, and that E-RWSC matches or outperform several baselines at substantially lower average cost (Section 4). For example, with Qwen-2.5-3B as the language model, ArmoRM as the reward model, and a maximum generation budget capped at 512, on the MATH500 Benchmark, we achieve an 8.2% absolute accuracy gain over BoN with a 61.7% smaller budget.

Related work. A line of work improves *which* response is chosen from a pool: reward-weighted selection with a temperature (Verdun et al., 2025), ensemble reward models normalized across members (Coste et al., 2024), and step-level search or process verification that refine how each candidate is generated and scored (Qiu et al., 2025; Chen et al., 2024; 2025; Lee et al., 2025a). RWSC draws on the first two but aggregates over final answers rather than individual candidates. Our contribution concerns the orthogonal question of *how many* candidates to generate. The adaptive stopping methods (Aggarwal et al., 2023; Wan et al., 2025; Kalayci et al., 2025; Raman et al., 2026; Sun et al., 2024) share a common limitation: their halting rules are tuned thresholds resting on distributional assumptions, with no control over the probability of stopping on the wrong answer. ConSol (Lee et al., 2025b), like SPRT-RWSC, applies Wald's test to halt self-consistency generation, but compares only the top two answers and uses static priors. Two recent works give anytime-valid certificates for self-consistency: MMC (Cordero-Encinar & Duncan, 2025) certifies the empirical-leader mode via paired e-processes and couples this with test-time training, and CITE (Ota et al., 2026) certifies a fixed target as the unique mode with a matching minimax lower bound.

2 Answer-Level Aggregation and Sequential Stopping

Given a prompt x, a language model samples responses from a fixed reference policy $\pi_{\mathrm{ref}}(\cdot|x)$. Each response y carries a chain of reasoning terminating in a final answer, extracted by a function $\mathrm{Ans}(y)$ that maps y to a canonical answer string (e.g., the final numerical result of a math problem). A reward model $r(x, y) \in \mathbb{R}$ scores how good a response is. ***Best-of-N*** (BoN) samples N responses and returns the one with the highest reward. ***Self-consistency*** ignores the reward model and returns the answer produced by the majority of responses. Scoring responses individually, as BoN does, exposes selection to reward hacking at its worst: a single anomalously scored response wins outright. Yet discarding the reward signal entirely, as self-consistency does, throws away information about response quality. Since many distinct reasoning chains reach the same final answer, we group responses at the answer level, scoring each answer group by both the number of responses it contains and the total reward it receives.

2.1 Reward-Weighted Self-Consistency

Let $\mathcal{Y}$ be a pool of N sampled responses, $\mathcal{Y}_a = \{y \in \mathcal{Y} : \mathrm{Ans}(y) = a\}$ the set of responses yielding answer a, and $\mathcal{A} = \{\mathrm{Ans}(y) : y \in \mathcal{Y}\}$ the set of observed answers. Reward-Weighted Self-Consistency (RWSC) assigns each group the softmax-weighted mass of its members' rewards,

$$S(x, a) \;=\; \sum_{y \in \mathcal{Y}_a} \frac{\exp\big(\overline{r}(x, y)/\tau\big)}{\sum_{y' \in \mathcal{Y}} \exp\big(\overline{r}(x, y')/\tau\big)}, \qquad \overline{r}(x, y) = \frac{r(x, y) - \mu_x}{\sigma_x}, \tag{1}$$

where $\overline{r}(x, y)$ is the reward normalized by the pool mean μ_x and standard deviation σ_x, and τ is a temperature parameter. The scores form a probability distribution over $\mathcal{A}$. As $\tau \to \infty$, every response contributes equally, and $S(x, a)$ reduces to the fraction of responses in cluster a, yielding self-consistency. As $\tau \to 0$, the mass concentrates on the highest-reward response, recovering BoN. Intermediate τ allows the answer frequency and reward magnitude to jointly determine the winner. RWSC returns the answer with the highest mass and, within its cluster, the highest-reward response:

$$a^* = \arg\max_{a \in \mathcal{A}} S(x, a), \qquad y^*_{\mathrm{RWSC}} = \arg\max_{y \in \mathcal{Y}_{a^*}} r(x, y). \tag{2}$$

RWSC, like BoN and self-consistency, draws a fixed N responses per prompt. This is wasteful, since the computation a prompt requires varies by orders of magnitude, and a budget large enough for the hardest prompt overspends on easy ones. Instead, we let each prompt govern its own budget, framing the decision of when to stop as a sequential test of the convergence of RWSC clusters.

2.2 A SEQUENTIAL TEST OVER ANSWER CLUSTERS

Fix a prompt x. Let the responses $y_1, y_2, \ldots$ be drawn sequentially from $\pi_{\text{ref}}(\cdot|x)$. At each step n, let $\mathcal{Y}_n$ be the pool of sampled responses, $\mathcal{A}_n$ the set of observed answers, $\mathcal{Y}_{n,a}$ the cluster for answer a and $S_n(x,a)$ its RWSC mass. Let $a_n^* = \arg\max_{a\in\mathcal{A}_n} S_n(x,a)$ be the current leader, the highest-RWSC-mass answer. For an answer a, write $\theta_a = \mathbb{P}_{y\sim\pi_{\text{ref}}(\cdot|x)}\big(\text{Ans}(y) = a\big)$ for its generation probability under π_{ref}. Deciding when to stop amounts to determining whether the leading answer's probability of generation is sufficiently high to declare it dominant. Each new response y_{n+1} contributes to an agreement indicator

$$X_{n+1} = \mathbb{1}\big\{\text{Ans}(y_{n+1}) = a_n^*\big\}, \quad \text{with} \quad \mathbb{P}\big(X_{n+1} = 1 \mid \mathcal{F}_{n-1}\big) = \theta_{a_n^*}, \tag{3}$$

where $\mathcal{F}_n = \sigma(\mathcal{Y}_n)$ denotes the sigma-algebra generated by first n responses and their reward scores. The test asks whether the agreement rate $\theta_{a_n^*}$ is high, so that the leader a_n^* has converged, or low, so that it has not. We instantiate it with Wald's sequential probability ratio test (SPRT) (Wald, 1945) and call the resulting procedure SPRT-RWSC. Its pseudo-code is given in Appendix A.

Hypothesis test. Fix $0 < p_0 < 0.5 < p_1 < 1$ and at each step n, we test two *simple* hypotheses,

$$H_0 : \theta_{a_n^*} = p_0 \quad \text{against} \quad H_1 : \theta_{a_n^*} = p_1, \tag{4}$$

where p_0 represents a non-convergent regime in which the leader holds only a minority of the mass and p_1 a regime in which it is dominant. The two rates are set per prompt from a short burn-in of n_0 responses (e.g. $n_0 = 4$), during which clusters are formed but no evidence is accumulated: we set $p_0 = \min\{1/|\mathcal{A}_{n_0}|, 0.5 - \epsilon\}$ and $p_1 = \max\{S_{n_0}(x, a_{n_0}^*),\ 0.5 + \epsilon\}$ for a small $\epsilon \in (0, 0.5)$, so that stopping always requires a strict majority. Under hypothesis H_k, $k \in \{0, 1\}$, each observation has mass $p_k^{X_i}(1-p_k)^{1-X_i}$, so the cumulative log-likelihood ratio after n responses,

$$\Lambda_n = \log\left(\prod_{i=n_0+1}^{n} \frac{p_1^{X_i}(1-p_1)^{1-X_i}}{p_0^{X_i}(1-p_0)^{1-X_i}}\right) = \sum_{i=n_0+1}^{n}\left[X_i \log\frac{p_1}{p_0} + (1-X_i)\log\frac{1-p_1}{1-p_0}\right], \tag{5}$$

accumulates evidence for H_1 over H_0. Given target error rates α (accepting a non-dominant leader) and β (rejecting a dominant one) in $(0, 1)$, Wald's boundaries are $A = \log\frac{\beta}{1-\alpha}$ and $B = \log\frac{1-\beta}{\alpha}$. If $\Lambda_n \geq B$, the test returns the current leader a_n^* as the final answer. If $\Lambda_n \leq A$, it rejects the leader and resets the evidence $\Lambda_n = 0$. Otherwise, it draws another response, falling back to the current leader if the budget $N_{\max}$ is reached.

Shifting leader. The RWSC leader a_n^* is a function of the responses and can change as we draw y_{n+1}. If the leader changes to $a_{n+1}^* \neq a_n^*$, the hypotheses – defined relative to the previous leader – no longer concern the quantity being tested, and the accumulated Λ_n measures evidence about the wrong answer. SPRT-RWSC resets $\Lambda_n = 0$ and re-estimates p_0, p_1 for the new leader with updated answer set $\mathcal{A}_{n+1}$ and leader's RWSC mass $S_{n+1}(x, a_{n+1}^*)$. Thus, the hypotheses are repeatedly redefined based on the data, allowing the test to adapt to the prompt's difficulty.

However, one cannot control the errors here as one can with Wald's theorem. Wald's classical guarantee is for two *known* hypotheses fixed in advance, with a single terminal decision. Here, the leader shifts during the run, changing the hypotheses, and the evidence Λ_n is compared to the boundary after every response rather than at a single stopping point. SPRT-RWSC is thus a principled and effective heuristic, with tunable parameters α, β, with no guarantee on Type I and II error.

3 ANYTIME-VALID CERTIFIED STOPPING

Wald's guarantee is not designed for hypotheses that shift with the data or for a statistic inspected after every response. We therefore turn to a framework built for exactly these conditions. *E-processes* (Ramdas et al., 2023) retain their validity under optional stopping, and predictable data-dependent construction. By building the test on e-processes, we control the probability of ever certifying a non-dominant answer at a user-specified level, uniformly over time.

Given a prompt x, let $A_n = \text{Ans}(y_n)$ be the final answer of response y_n and $X_n^{(a)} = \mathbb{1}\{A_n = a\}$ the indicator that response n agrees with answer a. Let θ_a be the generation probability of answer a. Then, $X_n^{(a)}$, $n \geq 1$, are i.i.d. Bernoulli(θ_a). For each answer a, we test

$$H_0^{(a)} :\ \theta_a \leq p^\dagger \qquad \text{against} \qquad H_1^{(a)} :\ \theta_a > p^\dagger, \tag{6}$$

where $p^\dagger \in (0,1)$ is threshold controlling dominance of an answer. Under $H_0^{(a)}$, a is not the dominant answer. Hence, rejecting $H_0^{(a)}$ certifies a as dominant. When $p^\dagger \geq \frac{1}{2}$, at most one answer can do so. So the certified answer, if it exists, is unique.

Per-answer betting e-processes. Let $X_n^{(a)} = \mathbb{1}\{A_n = a\}$ be the indicator that response n agrees with answer a and $\mathcal{F}_n^{(a)} = \sigma(y_1, X_1^{(a)}, \ldots, y_n, X_n^{(a)})$ the sigma-algebra generated by the first n responses, their reward scores, and agreement indicators with answer a. Let $(\lambda_n^{(a)})_{n\geq 1}$ be a stochastic process predictable with respect to the filtration $(\mathcal{F}_n^{(a)})_{n\geq 1}$, i.e., $\lambda_n^{(a)}$ is $\mathcal{F}_{n-1}^{(a)}$-measurable. Let $\sigma_a = \min\{n : A_n = a\}$ the first appearance of answer a. Define, for $n \geq \sigma_a$, the stochastic process

$$E_n^{(a)} = \prod_{i=\sigma_a+1}^{n} \Big(1 + \lambda_i^{(a)}\big(X_i^{(a)} - p^\dagger\big)\Big), \quad E_{\sigma_a}^{(a)} = 1. \tag{7}$$

The above admits a betting interpretation (Ramdas et al., 2023): a bettor, starting with unit capital, bets repeatedly on the event that the next candidate agrees with the answer a, at odds set by the null $\theta_a \leq p^\dagger$. Staking a fraction $\lambda_i^{(a)} p^\dagger$ of current capital returns $\lambda_i^{(a)}(1-p^\dagger)$ on agreement and forfeits $\lambda_i^{(a)} p^\dagger$ otherwise, so $E_n^{(a)}$ is the bettor's wealth after n rounds.

Stopping rule. Let $J(a) \in \{1, 2, \ldots\}$ be the rank of answer a in order of first appearance. Assign weights $\pi_j = \frac{1}{j(j+1)}$, so that $\sum_{j\geq 1} \pi_j = 1$. Given a certification level $\alpha \in (0,1)$, we stop generating responses at

$$N = \min\big\{n \geq 1 : \exists\, a \text{ with } \pi_{J(a)}\, E_n^{(a)} \geq \tfrac{1}{\alpha}\big\} \wedge N_{\max}. \tag{8}$$

If some answer a^* triggers the boundary, it is *certified*, and the returned candidate is the highest-reward trajectory in its cluster, $y^* = \arg\max_{y \in \mathcal{Y}_{a^*}} r(x, y)$. If the budget $N_{\max}$ is exhausted first, the procedure falls back to uncertified RWSC selection (equation 2).

Theorem 1 (Anytime-valid certification)**.** *Fix $\alpha, p^\dagger, c \in (0,1)$. For any answer distribution $\{\theta_a\}_a$, any budget $N_{\max} < \infty$, and any predictable sequence of bets $(\lambda_i^{(a)})_{i,a} \subset [0,\, c/p^\dagger]$,*

$$\mathbb{P}\left(\exists\, n,\ \exists\, a \text{ with } \theta_a \leq p^\dagger : a \text{ is certified at time } n\right) \leq \alpha. \tag{9}$$

Equivalently, with probability greater than $1-\alpha$, uniformly over all n and over all answer a such that a is certified at time n, satisfies $\theta_a > p^\dagger$.

Proof sketch. Certification of answer a occurs when the bettor's wealth reaches $1/(\pi_{J(a)}\alpha)$. For an answer that is not dominant ($\theta_a \leq p^\dagger$), $E_n^{(a)}$ is a nonnegative supermartingale, so by Ville's inequality, it attains this level with probability at most $\pi_{J(a)}\alpha$. Summing the weights $\pi_{J(a)}$ over answers bounds the overall probability of ever certifying a non-dominant answer by α. Complete proof is given in Appendix B.

Remark 1. *The bets enter Theorem 1 only through their range and predictability; they may depend arbitrarily on past responses and rewards, so the reward model cannot affect the certificate's validity. A change in the RWSC leader is likewise immaterial: each answer accumulates evidence in its own e-process, so no statistic is invalidated or discarded when the leader shifts. Finally, Theorem 1 certifies* dominance under the policy, *$\theta_{a^*} > p^\dagger$, not correctness: a confidently mistaken policy is certified quickly on a wrong answer. What links certification to accuracy is the self-consistency premise that the majority vote is usually correct, which we verify empirically (Section 4).*

3.1 Reward-informed betting

We now turn to that efficiency: given the freedom to choose any predictable bet, which choice certifies the dominant answer fastest, and how does the reward model inform that choice? Since the wealth $E_n^{(a)}$ compounds multiplicatively, the natural efficiency criterion is the *growth rate* $\frac{1}{n}\log E_n^{(a)}$, and the bet maximizing it is the classical Kelly bet (Kelly, 1956; Breiman, 2011).

Lemma 1 (Kelly bet)**.** *Let $X \sim \text{Bernoulli}(\theta)$ and, for $\lambda \in [0, 1/p^\dagger)$, let $G_\theta(\lambda) = \mathbb{E}[\log(1 + \lambda(X - p^\dagger))]$ be the expected one-step log-increment in wealth. Then G_θ is strictly concave, and its unconstrained maximizer is the* Kelly bet *$\lambda^\star(\theta) = \frac{\theta - p^\dagger}{p^\dagger(1-p^\dagger)}$. When admissible, i.e. $\lambda^\star(\theta) < 1/p^\dagger$, it attains the growth rate $G_\theta(\lambda^\star(\theta)) = d_{\mathrm{KL}}(\theta \,\|\, p^\dagger)$.*

Applied to answer a, the numerator of the bet is the bettor's edge over null, and the denominator is the variance of evidence under null. So the optimal stake is proportional to the signal-to-noise ratio of the evidence, and it is non-positive exactly when $\theta_a \le p^\dagger$, where no favorable wager exists. Maximizing expected log wealth, rather than expected wealth, is necessary: expected wealth is maximized by the all-in bet $\lambda = 1/p^\dagger$, which collapses the wealth $E_n^{(a)}$ to zero on the first disagreement and cannot recover. Restricting to $\lambda \le c/p^\dagger$ with $c < 1$ rules this out: every step then multiplies the wealth by at least $1 - c > 0$, so a disagreement shrinks $E_n^{(a)}$ but never collapses it, and the Kelly bet $\lambda^\star(\theta_a)$ within this range maximizes the growth rate $G_{\theta_a}(\lambda)$. When the bet is admissible, its growth rate is the KL-divergence between true and null agreement rates: evidence accrues at a rate equal to the statistical distinguishability of the two hypotheses. It matches the classical efficiency of Wald's SPRT (Wald, 1945) while retaining validity under composite nulls and optional stopping.

Plug-in bets. The Kelly bet depends on the unknown generation probability θ_a and so cannot be played. Instead, we form a predictable estimate of θ_a from the responses and their rewards observed so far. At each step $n \ge \sigma_a + 1$, where σ_a is the first appearance of answer a, we play the bet

$$\lambda_n^{(a)} = \mathrm{clip}\left(\frac{\widehat{\theta}_{n-1}^{(a)} - p^\dagger}{p^\dagger(1-p^\dagger)},\ 0,\ \frac{c}{p^\dagger}\right), \quad \text{where} \quad \widehat{\theta}_{n-1}^{(a)} = \frac{Z_{n-1}^{(a)} + m\, S_{n-1}(x,a)}{(n-1-\sigma_a) + m}. \tag{10}$$

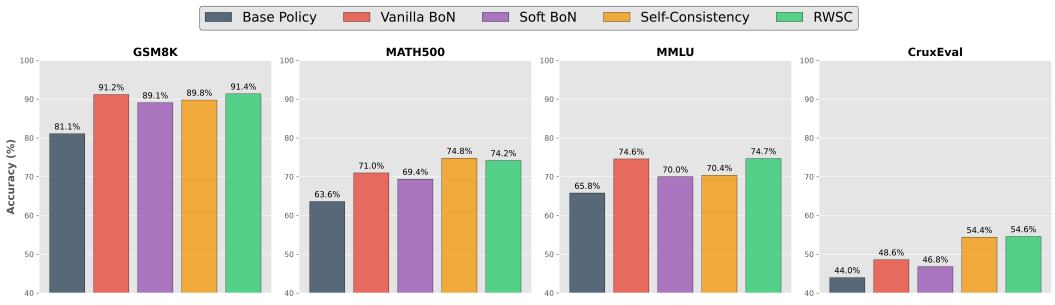
Here $Z_{n-1}^{(a)} = \sum_{j=\sigma_a+1}^{n-1} X_j^{(a)}$ denotes the number of agreements it accrued since its appearance, and hence the empirical agreement rate $Z_{n-1}^{(a)}/(n-1-\sigma_a)$ serves as an estimate of θ_a. We use a regularized estimator, where the current RWSC mass $S_{n-1}(x,a)$ of its cluster (equation 1) acts as a prior pseudo-observation of strength m (we use $m = 4$). The reward model enters the bet and, consequently, the test through the RWSC mass. We call this procedure E-RWSC. Its pseudo-code is given in Appendix A.

The clip keeps the bet in the admissible range $[0,\ c/p^\dagger]$. Informative reward scores steer $\widehat{\theta}_n^{(a)}$ toward θ_a early, sharpening the bets and accelerating certification; miscalibrated scores merely slow the growth of capital and cost efficiency, without affecting validity, since Theorem 1 holds for *any* predictable bet. Proposition 1 makes this precise, giving the asymptotic growth rate of the capital and the resulting certification time under the plug-in bet.

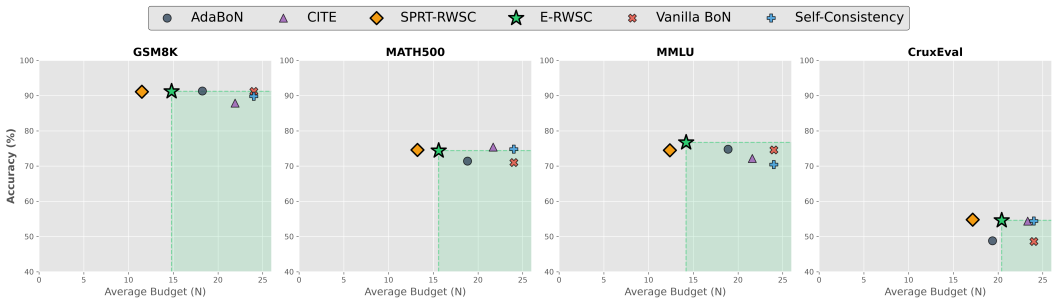
Proposition 1 (Growth rate and certification time)**.** *Let $\theta_{a^\star} > p^\dagger$ for an answer $a^\star$ and the Kelly bet $\lambda^\star(\theta_{a^\star}) < 1/p^\dagger$. Then, as $n \to \infty$, the plug-in bet $\lambda_n^{(a^\star)} \to \lambda^\star(\theta_{a^\star})$, and the capital $\frac{1}{n}\log E_n^{(a^\star)} \to d_{\mathrm{KL}}\big(\theta_{a^\star}||p^\dagger\big)$, almost surely. Let $N_{a^\star}(\alpha) = \min\{n : \pi_{J(a^\star)} E_n^{(a^\star)} \ge 1/\alpha\}$ be the time at which $a^\star$ is certified. Then $N_{a^\star}(\alpha) < \infty$ and as $\alpha \to 0$, $\frac{N_{a^\star}(\alpha)}{\log(1/\alpha)} \to \frac{1}{d_{\mathrm{KL}}(\theta_{a^\star}||p^\dagger)}$ almost surely. Moreover, when $p^\dagger \ge \frac{1}{2}$, the stopping time N of equation 8 coincides with $N_{a^\star}(\alpha)$ except on an event of probability at most α. consequently $N/\log(1/\alpha) \to 1/d_{\mathrm{KL}}(\theta_{a^\star}||p^\dagger)$ in probability as $\alpha \to 0$.*

Proposition 1 says that E-RWSC certifies a dominant answer in about $\log(1/\alpha)/d_{\mathrm{KL}}(\theta_{a^\star}||p^\dagger)$ responses: the cost grows only logarithmically in the confidence $1/\alpha$ and scales inversely with $d_{\mathrm{KL}}(\theta_{a^\star} \,\|\, p^\dagger)$, which measures how far the answer's mass sits above the threshold $p^\dagger$. Easy prompts, where one answer clearly dominates, are certified in a handful of responses; near-ties cost many, so computational cost adapts automatically to prompt difficulty. This rate is essentially the best attainable. Ota et al. (2026) proves a lower bound of the same order for sequential *mode* certification under self-consistency, where the sample complexity is governed by the margin between the target answer and its closest competitor. Instead, E-RWSC certifies that the returned answer holds an absolute majority, $\theta_a > \frac{1}{2}$ – stronger than being the mode – with governing quantity $d_{\mathrm{KL}}(\theta_a||\frac{1}{2})$. The rate transfers to the stopping time N of the test (equation 8) in probability. Proof is in Appendix B.

Remark 2. *Ota et al. (2026) certify self-consistency by an intersection-union test (*CITE*), with a confidence-weighted variant in which each vote carries a weight. Here, the weight enters the certified object, so a poor weight changes what is certified. In* E-RWSC*, the reward model enters only the bets, leaving the certified event $\theta_a > \frac{1}{2}$ fixed, so a poor reward can only slow certification, never invalidate it. More importantly,* CITE *fixes the target in advance, whereas* E-RWSC *runs per-answer e-processes and certifies whichever answer crosses first, uniformly over the growing answer set. Empirically,* E-RWSC *achieves higher accuracy at a lower average token cost than* CITE*.*



(a) Full Budget Baselines Accuracy



(b) Early-stopping Pareto Frontiers (Accuracy vs. Average Budget)

Figure 1: Comprehensive comparison of all selection methods across GSM8K, MATH500, MMLU and CruxEval for **Qwen2.5-3B-Instruct** using **ArmoRM** reward model, with allowed maximum budget as 24. For E-Process methods, we use $\alpha = 0.1$, $p^{\dagger} = 0.5$, and $c = 0.8$. For SPRT we use α=0.1, β=0.2.

4 Experiments

To evaluate the efficacy and generalizability of our proposed E-SPRT framework, we conduct experiments across four datasets. We utilize GSM8K (Cobbe et al., 2021) and MATH500 (Hendrycks et al., 2021b) to assess multi-step mathematical reasoning. To evaluate broad domain knowledge and academic reasoning, we benchmark on MMLU (Hendrycks et al., 2021a). We test the framework's capability in code reasoning and execution tracking using the CruxEval benchmark (Gu et al., 2024). The generation capabilities of our framework were analyzed across four distinct base language models: Qwen2.5-3B, Qwen2.5-7B, Llama-3-8B, and Phi-3-Mini-4K. To evaluate the generated reasoning paths, we utilized a diverse ensemble of four reward models: ArmoRM (Wang et al., 2024), Skywork V1 (Liu et al., 2024) and V2 (Liu et al., 2026), and FsfairX (Dong et al., 2024).

Comparison with Baselines. Figure 1 (a) compares RWSC against full-budget baselines and Figure 1 (b) compares our framework SPRT-RWSC and E-RWSC against early-stopping baselines with $N_{\max} = 24$. Across all evaluated tasks, RWSC consistently outperforms all full-budget baselines, achieving up to a 6% accuracy improvement over Vanilla BoN on CruxEval. Among sequential methods, AdaBoN (Raman et al., 2026) maintains baseline accuracy but fails to meaningfully reduce compute, consistently requiring $\sim$18$-$19 candidates on average. In contrast, SPRT-RWSC and E-RWSC successfully cut the required budget nearly in half ($\sim$11$-$17 candidates) while simultaneously matching or strictly exceeding the accuracy of full-budget baselines. As highlighted by the shaded dominance regions in Figure 1 (b), our methods, E-RWSC and SPRT-RWSC, consistently achieve the optimal Pareto frontier, outperforming CITE on both accuracy and budget reduction over 3 benchmarks and demonstrating a superior balance between alignment quality and computational efficiency. For a more detailed comparison with all the baselines, refer to Table 4 in the appendix.

Cross-Model Generalization and Reward Model Sensitivity. We conduct a two-dimensional ablation on MATH500. Table 1 applies a fixed reward model (ArmoRM) across four generators (3B to 8B). While absolute performance scales with base policy strength (peaking at 81.8% for Qwen2.5-7B), SPRT-RWSC and E-RWSC consistently dominate BoN accuracy across all models, cutting inference budgets by up to 54% (less on the harder-to-converge Llama-3-8B and Phi-3-Mini-4K). Moreover, the stopping rule adapts to generator confidence, expanding budgets for uncertain models (22.45 for Llama-3-8B) and halting early for highly convergent ones (13.06 for Qwen2.5-7B). Table 2 fixes the generator (Qwen2.5-3B) to test RM sensitivity. BoN is sensitive to RM calibration,

Method	Qwen2.5-3B		Qwen2.5-7B		Llama-3-8B		Phi-3-Mini-4K	
	Acc	**Budget**	**Acc**	**Budget**	**Acc**	**Budget**	**Acc**	**Budget**
π_{ref} (Base)	63.6%	1.00	75.6%	1.00	26.2%	1.00	39.2%	1.00
Vanilla BoN	71.0%	24.00	77.8%	24.00	32.6%	24.00	48.2%	24.00
RWSC	74.2%	24.00	81.4%	24.00	39.6%	24.00	56.4%	24.00
Pandora ($c = 0.05$)	68.6%	$4.00_{\pm 0.00}$	78.0%	$4.00_{\pm 0.00}$	30.0%	$4.00_{\pm 0.00}$	44.2%	$4.00_{\pm 0.00}$
AdaBoN	71.4%	$18.82_{\pm 5.20}$	78.0%	$18.97_{\pm 5.26}$	32.2%	$18.90_{\pm 5.31}$	47.8%	$18.84_{\pm 5.48}$
BEACON	70.8%	$22.55_{\pm 4.27}$	77.6%	$22.28_{\pm 4.70}$	32.0%	$22.39_{\pm 4.57}$	47.8%	$22.66_{\pm 4.12}$
ConSol	75.4%	$8.02_{\pm 7.38}$	80.8%	$5.91_{\pm 5.99}$	38.6%	$14.69_{\pm 8.38}$	53.4%	$12.02_{\pm 8.64}$
CITE ($\alpha = 0.1$)	75.4%	$21.68_{\pm 2.85}$	81.8%	$20.78_{\pm 2.92}$	39.0%	$23.74_{\pm 1.15}$	53.6%	$23.11_{\pm 2.02}$
Ours (Proposed)								
SPRT-RWSC ($n_0 = 4$)	74.6%	$13.23_{\pm 7.30}$	81.6%	$11.09_{\pm 6.46}$	39.6%	$19.33_{\pm 6.28}$	56.6%	$16.70_{\pm 7.26}$
E-RWSC ($\alpha = 0.1$)	74.4%	$15.59_{\pm 8.20}$	81.4%	$13.06_{\pm 7.81}$	39.6%	$22.45_{\pm 4.63}$	56.4%	$20.16_{\pm 6.75}$

Table 1: Cross-Model Performance Comparison on MATH500 ($N = 24$, ArmoRM).

Method	ArmoRM		Skywork V1		FsfairX		Skywork V2	
	Acc	**Budget**	**Acc**	**Budget**	**Acc**	**Budget**	**Acc**	**Budget**
π_{ref} (Base)	63.6%	1.00	63.6%	1.00	63.6%	1.00	63.6%	1.00
Vanilla BoN	71.0%	24.00	68.6%	24.00	62.6%	24.00	76.8%	24.00
RWSC	74.2%	24.00	76.0%	24.00	72.8%	24.00	78.8%	24.00
Pandora ($c = 0.05$)	68.6%	$4.00_{\pm 0.00}$	69.0%	$14.91_{\pm 8.27}$	64.0%	$6.11_{\pm 5.08}$	72.4%	$9.73_{\pm 7.74}$
AdaBoN	71.4%	$18.82_{\pm 5.20}$	69.0%	$19.59_{\pm 5.14}$	63.0%	$19.09_{\pm 5.29}$	75.8%	$19.10_{\pm 5.56}$
BEACON	70.8%	$22.55_{\pm 4.27}$	69.4%	$21.53_{\pm 5.65}$	62.6%	$22.23_{\pm 4.91}$	77.2%	$21.77_{\pm 5.41}$
ConSol	75.4%	$8.02_{\pm 7.38}$	75.4%	$8.02_{\pm 7.38}$	75.4%	$8.02_{\pm 7.38}$	75.4%	$8.02_{\pm 7.38}$
CITE ($\alpha = 0.1$)	75.4%	$21.68_{\pm 2.85}$	75.4%	$21.68_{\pm 2.85}$	75.4%	$21.68_{\pm 2.85}$	75.4%	$21.68_{\pm 2.85}$
Ours (Proposed)								
SPRT-RWSC ($n_0 = 4$)	74.6%	$13.23_{\pm 7.30}$	75.8%	$13.61_{\pm 7.54}$	72.6%	$13.51_{\pm 7.39}$	78.2%	$12.94_{\pm 7.17}$
E-RWSC ($\alpha = 0.1$)	74.4%	$15.59_{\pm 8.20}$	76.0%	$15.63_{\pm 8.19}$	72.6%	$15.65_{\pm 8.23}$	78.6%	$15.50_{\pm 8.18}$

Table 2: Cross-RM Sensitivity Analysis on MATH500 (Qwen2.5-3B, $N = 24$).

dropping to 62.6% with the weaker FsfairX RM and surging to 76.8% with Skywork V2. In contrast, RWSC and its sequential variants act as powerful regularizers against RM failures. E-RWSC maintains a robust 72.6% accuracy even with FsfairX and scales to 78.6% with Skywork V2.

4.1 Ablation Studies

Budget scaling and prevention of premature convergence. Table 3 illustrates the scaling behavior of selection methods as the maximum allowed budget expands from $N = 24$ to $N = 512$. A critical failure mode, *reward hacking*, of BoN is exposed here: accuracy degrades at high budgets (dropping from 71.4% at $N = 32$ to 69.8% at $N = 512$). RWSC successfully regularizes this effect, monotonically scaling accuracy up to 77.8%. ConSol saturates early (halting at $\sim$12 for $N_{\max} = 512$), resulting in stagnant accuracy even when the budget is much higher. In contrast, E-RWSC effectively bridges the gap between efficiency and inference scaling. It adaptively consumes more budget as $N_{\max}$ increases (from 15.6 to 196.1), selectively spending compute on highly uncertain prompts. Figure 2(a) shows that around 225 prompts out of 500 stop in less than 8 steps, whereas 174 extremely hard problems reach the maximum budget (detailed analysis is in Appendix D).

Sensitivity to e-process parameters (α and $p^\dagger$). The parameters α (Type-I error tolerance) and $p^\dagger$ (dominance threshold) govern the anytime-valid certification guarantee. Figure 2(b) demonstrates that E-RWSC is robust to α. The average budget decreases smoothly from 34.22 to 30.11 as α is relaxed from 0.05 to 0.4. This confirms that the practitioner can tune α to trade off statistical confidence for computational efficiency without a drop in accuracy. Figure 2(c) sweeps the threshold $p^\dagger$. Because $p^\dagger$ represents the hypothesized null-agreement rate, a higher $p^\dagger$ demands stronger empirical consensus before the plug-in Kelly bet yields positive capital growth. Consequently, the generation budget scales monotonically with $p^\dagger$ (from 15.56 at $p^\dagger = 0.1$ to 60.83 at $p^\dagger = 0.9$). Accuracies peak broadly in the middle ranges ($0.3 \leq p^\dagger \leq 0.8$), validating our default choice of $p^\dagger = 0.5$ as a sound threshold for certifying a strict majority.

Method	$N = 24$		$N = 32$		$N = 64$		$N = 128$		$N = 256$		$N = 512$	
	Acc	Budget	Acc	Budget	Acc	Budget	Acc	Budget	Acc	Budget	Acc	Budget
Vanilla BoN	71.0%	24.0	71.4%	32.0	69.2%	64.0	70.2%	128.0	69.4%	256.0	69.8%	512.0
RWSC	74.2%	24.0	76.0%	32.0	77.0%	64.0	77.6%	128.0	77.6%	256.0	77.8%	512.0
Pandora	68.6%	4.0	68.6%	4.0	68.6%	4.0	68.6%	4.0	68.6%	4.0	68.6%	4.0
AdaBoN	71.4%	$18.8_{\pm 5.2}$	70.6%	$24.8_{\pm 7.3}$	70.0%	$48.5_{\pm 15.6}$	70.2%	$95.9_{\pm 32.2}$	69.8%	$190.8_{\pm 65.5}$	69.4%	$380.5_{\pm 132.0}$
BEACON	70.8%	$22.6_{\pm 4.3}$	71.0%	$29.5_{\pm 6.7}$	69.0%	$56.8_{\pm 17.4}$	70.4%	$109.7_{\pm 40.4}$	69.6%	$212.9_{\pm 89.1}$	70.0%	$416.5_{\pm 190.0}$
CONSOL	75.4%	$8.0_{\pm 7.4}$	74.4%	$8.7_{\pm 9.0}$	74.6%	$10.1_{\pm 13.4}$	74.8%	$11.2_{\pm 19.0}$	74.8%	$11.9_{\pm 24.3}$	74.8%	$12.3_{\pm 29.3}$
CITE	75.4%	21.7±2.9	74.4%	26.1±6.6	75.6%	38.4±20.2	76.8%	56.5±45.9	77.2%	84.2±93.2	77.4%	124.9±175.9
SPRT-RWSC	74.6%	$13.2_{\pm 7.3}$	75.8%	$14.8_{\pm 9.8}$	76.2%	$17.5_{\pm 15.7}$	76.4%	$18.0_{\pm 17.3}$	76.4%	$18.1_{\pm 18.2}$	76.4%	$18.1_{\pm 18.2}$
E-RWSC	74.4%	$15.6_{\pm 8.2}$	76.0%	$19.3_{\pm 12.0}$	77.0%	$32.9_{\pm 27.1}$	77.8%	$57.9_{\pm 57.0}$	77.8%	$105.6_{\pm 117.1}$	78.0%	$196.1_{\pm 236.2}$

Table 3: MATH500 Budget Sweep Results (Qwen2.5-3B, ArmoRM)

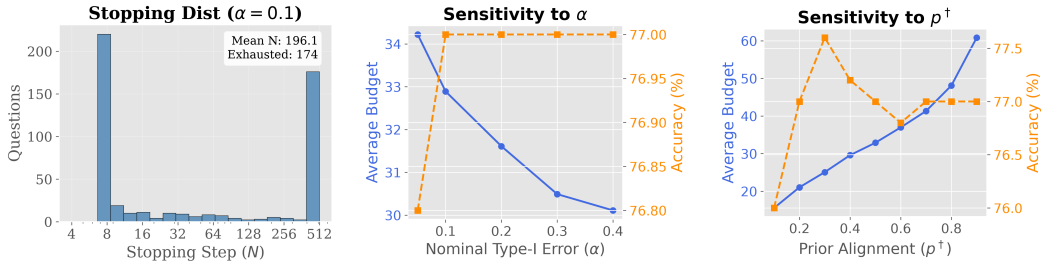


Figure 2: Ablation Studies on MATH500. (a) E-RWSC stopping distribution for $N_{\max} = 512$. (b) Sensitivity to Type-I error boundary α ($N_{\max} = 64$). (c) Sensitivity to the prior alignment threshold $p^\dagger$ ($N_{\max} = 64$).

Empirical validation of theoretical results.

We empirically validate E-RWSC theoretical guarantees in Figure 3. First, we validate the anytime-valid certification bound (Theorem 1) by sweeping the Type-I error rate α and measuring the realized error rate of E-RWSC. The realized error is defined as the fraction of prompts for which the sequential test halts early but certifies an answer that disagrees with the full-budget ($N = 512$) consensus. We see that the realized error rate is much less than the permitted rate α, indicating the validity and efficiency of early stopping.

Second, we validate Proposition 1, which governs the asymptotic growth rate of evidence. The theory states that the expected number of steps to stop scales with the inverse Kullback-Leibler divergence $1/d_{\mathrm{KL}}(\hat{\theta}\|p^\dagger)$ between the empirical answer distribution and the null hypothesis. As shown in the bottom panel, there is a linear relationship between the inverse of the signal strength and the stopping time N, confirming that our thresholding effectively controls the early-stopping budget, which is proportional to the prompt's inherent difficulty.

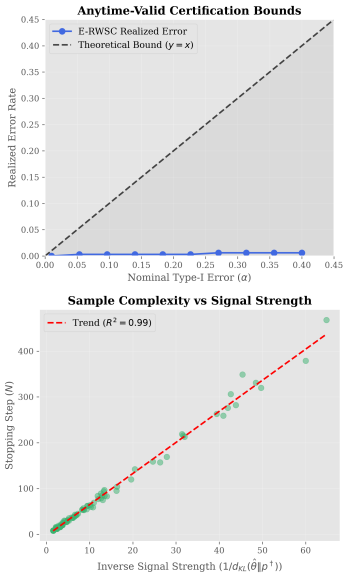


Figure 3: Empirical validation of E-RWSC theoretical bounds. (Top) Realized error rate respects the Type-I error (α) threshold. (Bottom) The inverse of signal strength versus stopping step shows a linear relationship.

5 Conclusion

We introduced Reward-Weighted Self-Consistency (RWSC), an answer-level aggregation rule spanning Best-of-N and self-consistency, and cast the question of how many candidates a prompt needs as a sequential test over its clusters: Wald's SPRT applied to RWSC works well empirically but cannot guarantee its error rate under a shifting leader, so we introduce with E-RWSC, which maintains a betting e-process per answer and bounds the probability of ever certifying a non-dominant answer by α for *any* predictable bets, at the minimax-optimal $O(\log(1/\alpha)/d_{\mathrm{KL}}(\theta_{a^\star}\|p^\dagger))$ rate under Kelly betting. Across four models, four reward models, and four benchmarks, e-RWSC matches or exceeds Best-of-N accuracy at substantially reduced cost (e.g., 8.2% higher accuracy than Vanilla BoN with 61.7% less budget on MATH500), with realized error and per-response cost tracking these bounds empirically.

AI USE STATEMENT

We have not used generative AI tools to generate synthetic data sets, help develop theoretical models or conceptual frameworks, assist with writing proofs, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis, interpret results. We have used generative AI tools to create or edit software code, summarize or analyze existing literature, source or search for information, and edit a paragraph to improve readability. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness. LLM-generated texts are also thoroughly checked and verified for correctness by every author. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not have any ethical concerns.

REPRODUCIBILITY STATEMENT

To ensure the full reproducibility of our empirical results, we provide our complete codebase as supplementary material. This repository contains the reference implementations for all proposed algorithms (including E-RWSC and SPRT-RWSC) and the evaluated baselines. Furthermore, we include the exact generation scripts used to interface with the language models, the data processing pipelines, and all helper functions necessary to extract answers and aggregate the final results. For our theoretical contributions, all underlying assumptions are explicitly stated in the main text, with complete proofs provided in Appendix B.

REFERENCES

Pranjal Aggarwal, Aman Madaan, Yiming Yang, et al. 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*, pp. 12375–12396, 2023.

Ahmad Beirami, Alekh Agarwal, Jonathan Berant, Alexander Nicholas D'Amour, Jacob Eisenstein, Chirag Nagpal, and Ananda Theertha Suresh. Theoretical guarantees on the best-of-n alignment policy. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=u3U8qzFV7w.

L Breiman. Optimal gambling systems for favorable games. *The Kelly Capital Growth Investment Criterion: Theory and Practice*, 3:47, 2011.

Guoxin Chen, Minpeng Liao, Chengxi Li, and Kai Fan. Alphamath almost zero: process supervision without process. *Advances in Neural Information Processing Systems*, 37:27689–27724, 2024.

Jiefeng Chen, Jie Ren, Xinyun Chen, Chengrun Yang, Ruoxi Sun, Jinsung Yoon, and Sercan O Arik. SETS: Leveraging self-verification and self-correction for improved test-time scaling. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=Wv9NMJoKww.

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

Paula Cordero-Encinar and Andrew B Duncan. Certified self-consistency: Statistical guarantees and test-time training for reliable reasoning in llms. *arXiv preprint arXiv:2510.17472*, 2025.

Thomas Coste, Usman Anwar, Robert Kirk, and David Krueger. Reward model ensembles help mitigate overoptimization. In *International Conference on Learning Representations*, volume 2024, pp. 50905–50931, 2024.

Hanze Dong, Wei Xiong, Bo Pang, Haoxiang Wang, Han Zhao, Yingbo Zhou, Nan Jiang, Doyen Sahoo, Caiming Xiong, and Tong Zhang. RLHF workflow: From reward modeling to online RLHF. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a13aYUU9eU.

Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

Alex Gu, Baptiste Roziere, Hugh James Leather, Armando Solar-Lezama, Gabriel Synnaeve, and Sida Wang. CRUXEval: A benchmark for code reasoning, understanding and execution. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 16568–16621. PMLR, 21–27 Jul 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. In J. Vanschoren and S. Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021b. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/be83ab3ecd0db773eb2dc1b0a17836a1-Paper-round2.pdf.

Yusuf Kalayci, Vinod Raman, and Shaddin Dughmi. Optimal stopping vs best-of-n for inference time optimization. *arXiv preprint arXiv:2510.01394*, 2025.

John L. Kelly. A new interpretation of information rate. *the bell system technical journal*, 35(4): 917–926, 1956.

Hyunseok Lee, Seunghyuk Oh, Jaehyung Kim, Jinwoo Shin, and Jihoon Tack. Revise: Learning to refine at test-time via intrinsic self-verification. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267, pp. 33616–33634, 2025a.

Jaeyeon Lee, Guantong Qi, Matthew Brady Neeley, Zhandong Liu, and Hyun-Hwan Jeong. Consol: Sequential probability ratio testing to find consistent llm reasoning paths efficiently. *arXiv preprint arXiv:2503.17587*, 2025b.

Chris Yuhao Liu, Liang Zeng, Jiacai Liu, Rui Yan, Jujie He, Chaojie Wang, Shuicheng Yan, Yang Liu, and Yahui Zhou. Skywork-reward: Bag of tricks for reward modeling in llms. *arXiv preprint arXiv:2410.18451*, 2024.

Yuhao Liu, Liang Zeng, Yuzhen Xiao, Jujie He, Jiacai Liu, Chaojie Wang, Rui Yan, Wei Shen, Fuxiang Zhang, Jiacheng Xu, et al. Skywork-reward-v2: Scaling preference data curation via human-ai synergy. In *International Conference on Learning Representations*, volume 2026, pp. 133805–133838, 2026.

Hirofumi Ota, Naoto Iwase, Yuki Ichihara, Junpei Komiyama, and Masaaki Imaizumi. Cite: Anytime-valid statistical inference in llm self-consistency. *arXiv preprint arXiv:2605.05873*, 2026.

Jiahao Qiu, Yifu Lu, Yifan Zeng, Jiacheng Guo, Jiayi Geng, Chenhao Zhu, Xinzhe Juan, Ling Yang, Huazheng Wang, Kaixuan Huang, et al. Treebon: Enhancing inference-time alignment with speculative tree-search and best-of-n sampling. In *EMNLP (Findings)*, pp. 20894–20917, 2025.

Vinod Raman, Hilal Asi, and Satyen Kale. Adabon: Adaptive best-of-n alignment, 2026. URL https://arxiv.org/abs/2505.12050.

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

Hanshi Sun, Momin Haider, Ruiqi Zhang, Huitao Yang, Jiahao Qiu, Ming Yin, Mengdi Wang, Peter Bartlett, and Andrea Zanette. Fast best-of-n decoding via speculative rejection. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=348hfcprUs.

Claudio Mayrink Verdun, Alex Oesterling, Himabindu Lakkaraju, and Flavio P Calmon. Soft best-of-n sampling for model alignment. In *2025 IEEE International Symposium on Information Theory (ISIT)*, pp. 1–6. IEEE, 2025.

Jean Ville. *Étude critique de la notion de collectif*, volume 3. Gauthier-Villars Paris, 1939.

A. Wald. Sequential Tests of Statistical Hypotheses. *The Annals of Mathematical Statistics*, 16(2): 117 – 186, 1945. doi: 10.1214/aoms/1177731118.

Guangya Wan, Zixin Stephen Xu, Sasa Zorc, Manel Baucells, Mengxuan Hu, Hao Wang, and Sheng Li. Beacon: Bayesian optimal stopping for efficient llm sampling. *arXiv preprint arXiv:2510.15945*, 2025.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10582–10592, 2024.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models, 2022.

A PSEUDOCODES

Algorithm 1 SPRT-RWSC

Require: prompt x; policy π_{ref}; reward model r; errors α, β; budget $N_{\max}$; burn-in n_0

$A \leftarrow \log\left(\frac{\beta}{1-\alpha}\right)$; $\quad B \leftarrow \log\left(\frac{1-\beta}{\alpha}\right)$
Generate candidates $y_1, \ldots, y_{n_0} \sim \pi_{\text{ref}}(\cdot \mid x)$; $\quad \mathcal{A} \leftarrow \{\text{Ans}(y_i)\}_{i=1}^{n_0}$ ▷ Burn-in phase
Construct initial answer clusters and RWSC masses $S_{n_0}(x, \cdot)$
$a^* \leftarrow \arg\max_{a \in \mathcal{A}} S_{n_0}(x, a)$ ▷ Identify initial leader
$\Lambda \leftarrow 0$; $\quad$ set p_0, p_1 ▷ Initialize Log-Likelihood Ratio
for $n = n_0 + 1, \ldots, N_{\max}$ **do**
    Sample $y_n \sim \pi_{\text{ref}}(\cdot \mid x)$; $\quad A_n \leftarrow \text{Ans}(y_n)$; $\quad \mathcal{A} \leftarrow \mathcal{A} \cup \{A_n\}$
    Update answer clusters and RWSC masses $S_n(x, \cdot)$ over $y_{1:n}$
    $a_{\text{new}} \leftarrow \arg\max_{a \in \mathcal{A}} S_n(x, a)$ ▷ Identify current leader
    if $a_{\text{new}} \neq a^*$ **then** ▷ Leader Shift
        $a^* \leftarrow a_{\text{new}}$; $\quad \Lambda \leftarrow 0$; $\quad$ Update p_0, p_1
    else
        $X \leftarrow \mathbf{1}\{A_n = a^*\}$ ▷ Check for agreement
        $\Delta\Lambda \leftarrow X \log\left(\frac{p_1}{p_0}\right) + (1 - X) \log\left(\frac{1-p_1}{1-p_0}\right)$; $\quad \Lambda \leftarrow \Lambda + \Delta\Lambda$
        if $\Lambda \geq B$ **then** ▷ Accept H_1
            return a^* ▷ Stop early and return leader
        else if $\Lambda \leq A$ **then** ▷ Reject H_1
            $\Lambda \leftarrow 0$; $\quad$ Update p_0, p_1 ▷ Reset evidence and track new trend
        end if
    end if
end for
return a^* ▷ Budget exhausted, return current leader

Algorithm 2 E-RWSC

Require: prompt x; policy π_{ref}; reward model r; level α; threshold $p^\dagger$ (default $\frac{1}{2}$); clip $c \in (0, 1)$; prior strength m; budget $N_{\max}$; weights $\pi_j = \frac{1}{j(j+1)}$

$\mathcal{A} \leftarrow \emptyset$; $\quad J \leftarrow 0$ ▷ active answers; count of distinct answers seen
for $n = 1, 2, \ldots, N_{\max}$ **do**
    for each $a \in \mathcal{A}$ **do** ▷ bets are predictable: computed *before* y_n is drawn
        $\hat{\theta}^{(a)} \leftarrow \dfrac{Z^{(a)} + m\, S_{n-1}(x, a)}{(n - 1 - \sigma_a) + m}$ ▷ RWSC mass $\widehat{S}$ from Eq. 1 on $y_{1:n-1}$
        $\lambda^{(a)} \leftarrow \text{clip}\Big(\frac{\hat{\theta}^{(a)} - p^\dagger}{p^\dagger(1-p^\dagger)},\, 0,\, \frac{c}{p^\dagger}\Big)$ ▷ plug-in Kelly bet, Eq. 10
    end for
    sample $y_n \sim \pi_{\text{ref}}(\cdot \mid x)$; $\quad$ score $r(x, y_n)$; $\quad A_n \leftarrow \text{Ans}(y_n)$
    recompute RWSC masses $S_n(x, \cdot)$ over $y_{1:n}$
    for each $a \in \mathcal{A}$ **do** ▷ settle all active bets against the new observation
        $X \leftarrow \mathbf{1}\{A_n = a\}$; $\quad E^{(a)} \leftarrow E^{(a)} \cdot \big(1 + \lambda^{(a)}(X - p^\dagger)\big)$; $\quad Z^{(a)} \leftarrow Z^{(a)} + X$
    end for
    if $A_n \notin \mathcal{A}$ **then** ▷ lazy instantiation; first observation consumed, not bet on
        $\mathcal{A} \leftarrow \mathcal{A} \cup \{A_n\}$; $\quad J \leftarrow J + 1$; $\quad J(A_n) \leftarrow J$
        $\sigma_{A_n} \leftarrow n$; $\quad E^{(A_n)} \leftarrow 1$; $\quad Z^{(A_n)} \leftarrow 0$
    end if
    if $\exists\, a \in \mathcal{A} : \pi_{J(a)}\, E^{(a)} \geq 1/\alpha$ **then** ▷ Ville boundary, Eq. 8
        $a^* \leftarrow \arg\max_{a \in \mathcal{A}} \pi_{J(a)} E^{(a)}$
        return $y^* \leftarrow \arg\max_{y \in \mathcal{Y}_{a^*}} r(x, y)$ ▷ **certified:** $\theta_{a^*} > p^\dagger$ w.p. $\geq 1 - \alpha$ (Thm. 1)
    end if
end for
return y^*_{RWSC} from equation 2 ▷ budget exhausted: uncertified fallback

B DEFERRED PROOFS

Proof of Theorem 1. Let $\mathcal{N} = \{a : \theta_a \le p^\dagger\}$ denote the set of null answers, and for each answer a let

$$C_a \;=\; \big\{\exists\, n \ge \sigma_a : \pi_{J(a)} E_n^{(a)} \ge 1/\alpha\big\} \tag{11}$$

be the event that a is certified at some time. Note that $C_a \subseteq \{\sigma_a < \infty\}$, since $E^{(a)}$ is instantiated only once a has been observed. We must bound $\mathbb{P}\big(\bigcup_{a\in\mathcal{N}} C_a\big)$.

Fix $a \in \mathcal{N}$ and work on the event $\{\sigma_a < \infty\}$, since outside this event there is no supermartingale — the process was never started, and $J(a)$ isn't even defined. The candidates are i.i.d. draws from $\pi_{\mathrm{ref}}(\cdot \mid x)$ and σ_a is a stopping time, so conditionally on $\mathcal{F}_{\sigma_a}$ the post-σ_a observations $\big(X_i^{(a)}\big)_{i>\sigma_a}$ are i.i.d. $\mathrm{Bernoulli}(\theta_a)$ and independent of $\mathcal{F}_{\sigma_a}$; starting the product at $\sigma_a + 1$ is what excludes the deterministic success $X_{\sigma_a}^{(a)} = 1$ implied by the definition of σ_a. Each factor $U_i = 1+\lambda_i^{(a)}(X_i^{(a)}-p^\dagger)$ is bounded below by $1 - c > 0$, since worst case $X_i^{(a)} = 0$ gives $1 - \lambda_i^{(a)} p^\dagger$ and $\lambda_i^{(a)} \le c/p^\dagger$, and the bets are $\mathcal{F}_{i-1}$-measurable, so

$$\mathbb{E}\big[U_i \,\big|\, \mathcal{F}_{i-1}\big] \;=\; 1 + \lambda_i^{(a)}\big(\theta_a - p^\dagger\big) \;\le\; 1, \tag{12}$$

using $\lambda_i^{(a)} \ge 0$ and $\theta_a \le p^\dagger$. Hence, under the conditional law given $\mathcal{F}_{\sigma_a}$, the process $n \mapsto (E_n^{(a)})_{n\ge\sigma_a}$ is a nonnegative supermartingale started at 1.

Ville's inequality bounds $\mathbb{P}(\sup_n M_n \ge u) \le 1/u$ for any nonnegative supermartingale with $M_0 = 1$. The rank $J(a)$ is determined by the answers observed up to and including time σ_a, hence is $\mathcal{F}_{\sigma_a}$-measurable; applying Ville at threshold $u = 1/(\pi_{J(a)}\alpha)$ gives

$$\mathbb{P}\big(C_a \,\big|\, \mathcal{F}_{\sigma_a}\big) = \mathbb{P}\Big(\sup_{n\ge\sigma_a} E_n^{(a)} \ge \tfrac{1}{\pi_{J(a)}\alpha} \,\Big|\, \mathcal{F}_{\sigma_a}\Big) \;\le\; \pi_{J(a)}\,\alpha \qquad \text{on } \{\sigma_a < \infty\}. \tag{13}$$

Since $C_a \subseteq \{\sigma_a < \infty\}$ and $\{\sigma_a < \infty\} \in \mathcal{F}_{\sigma_a}$, the tower property gives

$$\begin{aligned}\mathbb{P}(C_a) = \mathbb{E}\big[\mathbb{E}\big[\mathbf{1}_{C_a} \mid \mathcal{F}_{\sigma_a}\big]\big] = \mathbb{E}\big[\mathbb{E}\big[\mathbf{1}_{C_a}\mathbf{1}\{\sigma_a < \infty\} \mid \mathcal{F}_{\sigma_a}\big]\big] &= \mathbb{E}\big[\mathbb{P}(C_a \mid \mathcal{F}_{\sigma_a})\,\mathbf{1}\{\sigma_a < \infty\}\big] \\ &\le \alpha\,\mathbb{E}\big[\pi_{J(a)}\,\mathbf{1}\{\sigma_a < \infty\}\big]. \end{aligned} \tag{14}$$

The set of possible answer strings is countable, so countable subadditivity applies. Combining with equation 14 and exchanging sum and expectation by Tonelli's theorem (all terms nonnegative),

$$\mathbb{P}\Big(\bigcup_{a\in\mathcal{N}} C_a\Big) \;\le\; \sum_{a\in\mathcal{N}} \mathbb{P}(C_a) \;\le\; \alpha\,\mathbb{E}\Big[\sum_{a\in\mathcal{N},\,\sigma_a<\infty} \pi_{J(a)}\Big] \;\le\; \alpha\,\mathbb{E}\Big[\sum_{a:\,\sigma_a<\infty} \pi_{J(a)}\Big], \tag{15}$$

the last inequality by dropping the restriction to $\mathcal{N}$, which only adds nonnegative terms.

Fix a sample path. Answers are enumerated in order of first appearance, so $a \mapsto J(a)$ is injective on the observed answers $\{a : \sigma_a < \infty\}$; the ranks are therefore distinct positive integers and

$$\sum_{a:\,\sigma_a<\infty} \pi_{J(a)} \;\le\; \sum_{j\ge1} \pi_j \;=\; \sum_{j\ge1}\Big(\frac{1}{j} - \frac{1}{j+1}\Big) \;=\; 1. \tag{16}$$

This holds for *every* sample path, not merely in expectation, so the random variable inside the expectation in equation 15 is bounded above by the constant 1; hence its expectation is at most 1.

Substituting equation 16 into equation 15 gives $\mathbb{P}\big(\bigcup_{a\in\mathcal{N}} C_a\big) \le \alpha$. Truncation at $N_{\max}$ only removes elements from each C_a, so the bound holds for any budget; and since Ville's inequality controls the supremum over all times, the guarantee is unaffected by when the analyst chooses to stop, which is the anytime-validity property. □

Proof of Lemma 1. Obseve that $G_\theta(\lambda) = \mathbb{E}[\log(1 + \lambda(X - p^\dagger))] = \theta\log(1 + \lambda(1 - p^\dagger)) + (1 - \theta)\log(1 - \lambda p^\dagger)$. Strict concavity is immediate since $G''_\theta(\lambda) < 0$ on the admissible interval (a sum of two negative terms). Setting $G'_\theta(\lambda) = 0$ gives $\theta(1-p^\dagger)/(1+\lambda(1-p^\dagger)) = (1-\theta)p^\dagger/(1-\lambda p^\dagger)$; cross-multiplying and cancelling the terms linear in λ yields $\theta(1-p^\dagger) - \lambda p^\dagger(1-p^\dagger) = (1-\theta)p^\dagger$, which gives the bet. Substituting $\lambda^\star(\theta)$ into G_θ and simplifying, using $1 + \lambda^\star(1-p^\dagger) = \theta/p^\dagger$ and $1 - \lambda^\star p^\dagger = (1-\theta)/(1-p^\dagger)$, gives the growth rate. □

Proof of Proposition 1. Write $\theta = \theta_{a^*}$. Throughout, "$\to$" denotes almost-sure convergence unless stated otherwise, and all statements are made on the event $\{\sigma_{a^\star} < \infty\}$, which has probability one since $\theta > 0$. Recall the expected one-step log-increment

$$G_\theta(\lambda) = \mathbb{E}_{X\sim\mathrm{Ber}(\theta)}[\log(1+\lambda(X-p^\dagger))] = \theta\log(1+\lambda(1-p^\dagger)) + (1-\theta)\log(1-\lambda p^\dagger).$$

Abbreviate $\lambda_i = \lambda_i^{(a^\star)}, X_i = X_i^{(a^\star)}, \xi_i = \log(1+\lambda_i(X_i-p^\dagger))$, so that $\log E_n^{(a^\star)} = \sum_{i=\sigma_{a^\star}+1}^{n}\xi_i$.

(i) The plug-in bets converge. Conditionally on $\mathcal{F}_{\sigma_{a^\star}}$, the post-instantiation indicators $(X_i)_{i>\sigma_{a^\star}}$ are i.i.d. $\mathrm{Bernoulli}(\theta)$ (as in the proof of Theorem 1). By strong law of large numbers, the empirical agreement rate $\frac{Z_{i-1}^{(a^\star)}}{i-1-\sigma_{a^\star}} \to \theta$. The plug-in estimate (equation 10) is $\hat\theta_{i-1}^{(a^\star)} = \frac{Z_{i-1}^{(a^\star)} + m\,S_{i-1}(x,a^\star)}{(i-1-\sigma_{a^\star})+m}$, a convex combination of the empirical rate and the bounded prior (RWSC mass) $S_{i-1}(x,a^\star)\in[0,1]$. The prior weight $\frac{m}{(i-1-\sigma_{a^\star})+m} \to 0$. Hence $\hat\theta_{i-1}^{(a^\star)} \to \theta$ regardless of the prior sequence. The Kelly map $\theta' \mapsto \mathrm{clip}(\lambda^\star(\theta'), 0, c/p^\dagger)$ is continuous (a clip of the continuous map $\lambda^\star$), so by the continuous mapping theorem, the plug-in bet $\lambda_i \to \lambda^\star(\theta)$ when $\lambda^\star(\theta) \le c/p^\dagger$.

(ii) Convergence of captial. The bet λ_i is $\mathcal{F}_{i-1}$-measurable by construction, and X_i is independent of $\mathcal{F}_{i-1}$ with mean θ, so $\mathbb{E}[\xi_i \mid \mathcal{F}_{i-1}] = G_\theta(\lambda_i)$. The map G_θ is continuous on the compact interval $[0, c/p^\dagger]$ (a sum of logarithms with arguments bounded below by $1-c>0$), so by continuity and Step (i), $G_\theta(\lambda_i) \to G_\theta(\lambda^\star(\theta))$. The Cesàro lemma (Durrett, 2019) gives

$$\frac{1}{n}\sum_{i=\sigma_{a^\star}+1}^{n}\mathbb{E}[\xi_i \mid \mathcal{F}_{i-1}] \;\to\; G_\theta(\lambda^\star(\theta)). \tag{17}$$

Let $D_i = \xi_i - \mathbb{E}[\xi_i \mid \mathcal{F}_{i-1}]$, so that $(D_i)_i$ is a martingale-difference sequence with respect to $(\mathcal{F}_i)_i$. Because $\lambda_i \in [0, c/p^\dagger]$, each ξ_i lies in the bounded interval $[\log(1-c), \log(1+c(1-p^\dagger)/p^\dagger)]$, so $|D_i| \le L$ for a constant $L<\infty$; in particular $\sum_i \mathrm{Var}(D_i)/i^2 \le \sum_i L^2/i^2 < \infty$. Then, the martingale strong law (Durrett, 2019) yields $\frac{1}{n}\sum_{i\le n} D_i \to 0$. Adding this with equation 17, and noting the finitely many terms with index at most $\sigma_{a^\star}$ contribute $o(n)$ and so do not affect the Cesàro limit, gives $\frac{1}{n}\log E_n^{(a^\star)} \to G_\theta(\lambda^*(\theta)) = d_{\mathrm{KL}}(\theta\,\|\,p^\dagger)$ whenever $\lambda^\star(\theta) \le c/p^\dagger$.

(iii) Certification time. Fix a sample path in the probability-one event $\{\sigma_{a^\star} < \infty\}$, on which $\frac{1}{n}\log E_n^{(a^\star)} \to d_{\mathrm{KL}}(\theta\,\|\,p^\dagger)$. Since $\theta > p^\dagger$, we have $d_{\mathrm{KL}}(\theta\,\|\,p^\dagger) > 0$ and hence $\log E_n^{(a^\star)} \to \infty$. Write $b_\alpha = \log(1/(\pi_{J(a^\star)}\alpha))$ for the log-boundary of $a^\star$. Note $J(a^\star)$, hence $\pi_{J(a^\star)}$, is fixed on this path and does not depend on α, and $b_\alpha \to \infty$ as $\alpha \to 0$. Hence $N_{a^\star}(\alpha) = \min\{n : \log E_n^{(a^\star)} \ge b_\alpha\}$ is finite for every α since the walk diverges, and $N_{a^\star}(\alpha) \to \infty$ as $\alpha \to 0$. Now, by definition,

$$\log E_{N_{a^\star}(\alpha)-1}^{(a^\star)} \;<\; b_\alpha \;\le\; \log E_{N_{a^\star}(\alpha)}^{(a^\star)}. \tag{18}$$

On the fixed path, the sequence $m \mapsto \frac{1}{m}\log E_m^{(a^\star)}$ converges to $d_{\mathrm{KL}}(\theta\,\|\,p^\dagger)$, so along the divergent integer sequence $m = N_{a^\star}(\alpha)$ (and likewise $m = N_{a^\star}(\alpha)-1$, using $\frac{N_{a^\star}(\alpha)-1}{N_{a^\star}(\alpha)} \to 1$) both outer terms of equation 18, divided by $N_{a^\star}(\alpha)$, tend to $d_{\mathrm{KL}}(\theta\,\|\,p^\dagger)$. Hence $b_\alpha/N_{a^\star}(\alpha) \to d_{\mathrm{KL}}(\theta\,\|\,p^\dagger)$, i.e. $N_{a^\star}(\alpha)/b_\alpha \to d_{\mathrm{KL}}(\theta\,\|\,p^\dagger)$. Since $b_\alpha/\log(1/\alpha) = 1 + \log(1/\pi_{J(a^\star)})/\log(1/\alpha) \to 1$, we may replace b_α by $\log(1/\alpha)$, yielding $N_{a^\star}(\alpha)/\log(1/\alpha) \to 1/d_{\mathrm{KL}}(\theta\,\|\,p^\dagger)$ a.s.

(iv) Stopping time. N of equation 8 halts at the first answer to cross its boundary, so $N \le N_{a^\star}(\alpha)$ always, with equality when no other answer is certified first. When $p^\dagger \ge \frac{1}{2}$, $a^\star$ is the unique answer with $\theta_a > p^\dagger$, so any other certified answer is non-dominant. By Theorem 1, this occurs with probability at most α. Thus $\mathbb{P}(N \ne N_{a^\star}(\alpha)) \le \alpha$. Fix $\varepsilon > 0$. Since $N_{a^\star}(\alpha)/\log(1/\alpha) \to 1/d_{\mathrm{KL}}(\theta\,\|\,p^\dagger)$ almost surely, and hence in probability. So

$$\mathbb{P}\big(|N_{a^\star}(\alpha)/\log(1/\alpha) - 1/d_{\mathrm{KL}}(\theta\,\|\,p^\dagger)| > \varepsilon\big) \to 0.$$

Combined with $\mathbb{P}(N \ne N_{a^\star}(\alpha)) \le \alpha \to 0$, a union bound gives

$$\mathbb{P}\big(|N/\log(1/\alpha) - 1/d_{\mathrm{KL}}(\theta\,\|\,p^\dagger)| > \varepsilon\big) \to 0,$$

which completes the proof. □

C COMPARISION WITH BASELINES

Table 4 reports the full breakdown underlying the summary in Section 4. The Oracle@24 row upper-bounds what any selection rule can achieve from this candidate pool, and the gap to it is largest on CruxEval (76.5% vs. our best 54.8%) and smallest on GSM8K (96.5% vs. 91.4%), reflecting that code-execution answers are harder to concentrate reward mass on than short numeric ones. The base policy π_{ref} alone already reaches 81.1% on GSM8K without any selection, so the relative gains from BoN-style methods are naturally compressed there; on MATH500, MMLU, and CruxEval, where the base policy starts much lower (63.6%, 65.8%, 44.0%), selection contributes proportionally more of the final accuracy. Soft BoN (Verdun et al., 2025) underperforms Vanilla BoN on every benchmark, indicating that softening the arg-max selection with a temperature is not by itself sufficient to recover the consensus robustness that RWSC's explicit answer-level aggregation provides. Among the baselines, BEACON (Wan et al., 2025) takes a very high budget, where as for others, budget usage is highly Reward Model dependent or benchmark-dependent rather than a fixed property of the method: Pandora's Box BoN (Kalayci et al., 2025) is highly dependent on the scale of output of reward modes, with very confident answers of ArmoRM, it stops at budget same as burnin phase, although it changes as we change the reward model, as it can be seen in Table 2, ConSol's budget ranges from 5.39 on MMLU to 11.99 on CruxEval, and CITE's stays comparatively flat (21.60–23.31) across all four, consistent with its fixed-target certification not adapting its threshold to how concentrated a given benchmark's answer distribution is. SPRT-RWSC and E-RWSC track prompt difficulty instead: both spend the most on CruxEval (17.17 and 20.40, respectively) and the least on GSM8K (11.48 and 14.81), mirroring the same base-policy-strength ordering that governs the oracle gap.

	GSM8K		MATH500		MMLU		CruxEval	
Method	**Acc**	**Budget**	**Acc**	**Budget**	**Acc**	**Budget**	**Acc**	**Budget**
Oracle@24 (Hard Ceiling)	**96.5%**	24.00	**90.2%**	24.00	**93.4%**	24.00	**76.5%**	24.00
Full Budget Baselines								
Base Policy (π_{ref})	81.1%	24.00	63.6%	24.00	65.8%	24.00	44.0%	24.00
Vanilla BoN	91.2%	24.00	71.0%	24.00	74.6%	24.00	48.6%	24.00
Soft BoN	89.1%	24.00	69.4%	24.00	70.0%	24.00	46.8%	24.00
Answer-Level Aggregation								
Self-Consistency	89.8%	24.00	74.8%	24.00	70.4%	24.00	54.4%	24.00
RWSC	91.4%	24.00	74.2%	24.00	74.7%	24.00	54.6%	24.00
Early-stopping Baselines								
Pandora's Box BoN	88.1%	$4.00_{\pm 0.00}$	68.6%	$4.00_{\pm 0.00}$	73.4%	$4.00_{\pm 0.00}$	49.4%	$4.00_{\pm 0.00}$
AdaBoN	91.3%	$18.25_{\pm 5.90}$	71.4%	$18.82_{\pm 5.20}$	74.8%	$18.89_{\pm 5.48}$	48.8%	$19.37_{\pm 5.24}$
BEACON	91.3%	$22.06_{\pm 5.16}$	70.8%	$22.55_{\pm 4.27}$	74.6%	$22.26_{\pm 4.74}$	48.5%	$23.16_{\pm 3.22}$
ConSol	88.0%	$5.53_{\pm 4.81}$	75.4%	$8.02_{\pm 7.38}$	71.8%	$5.39_{\pm 4.56}$	54.1%	$11.99_{\pm 8.50}$
CITE	87.9%	$21.91_{\pm 2.72}$	75.4%	$21.68_{\pm 2.85}$	72.2%	$21.60_{\pm 2.85}$	54.4%	$23.31_{\pm 1.81}$
Proposed Methods								
SPRT-RWSC	91.1%	$11.48_{\pm 6.11}$	74.6%	$13.23_{\pm 7.30}$	74.5%	$12.38_{\pm 6.97}$	54.8%	$17.17_{\pm 7.05}$
E-RWSC	91.2%	$14.81_{\pm 7.82}$	74.4%	$15.59_{\pm 8.20}$	76.7%	$14.19_{\pm 7.86}$	54.6%	$20.40_{\pm 6.54}$

Table 4: Comprehensive comparison of all selection methods across GSM8K, MATH500, MMLU and CruxEval for **Qwen2.5-3B-Instruct** using **ArmoRM** reward model, with allowed maximum budget as 24. For E-Process methods, we use $\alpha = 0.1$, $p^{\dagger} = 0.5$, and $c = 0.8$. For SPRT we use α=0.1, β=0.2.

D STOPPING DISTRIBUTION ANALYSIS

To further understand the behavior of our anytime-valid stopping rule, we analyze the distribution of stopping times on the MATH500 dataset under a maximum budget of $N_{\max} = 512$ (Figure 4). The test naturally segments the dataset into "confident" queries, where the statistical boundary is reached early, and "highly uncertain" queries, where the framework exhausts the full budget (174 out of 500 questions). Figure 5 illustrates a stark performance divergence between these two subsets. On confident questions, the proxy reward model provides a strong, reliable signal, allowing both Vanilla BoN and our RWSC to perform efficiently. However, on the subset of highly uncertain questions, the reward model completely collapses, causing Vanilla BoN accuracy to plummet to 20.1%. In these failure modes, pure Self-Consistency remains significantly more robust (28.2%). Because our

framework dynamically falls back to broad topological consensus when proxy confidence is low, it successfully avoids the severe degradation of the reward model that traps standard parallel methods.

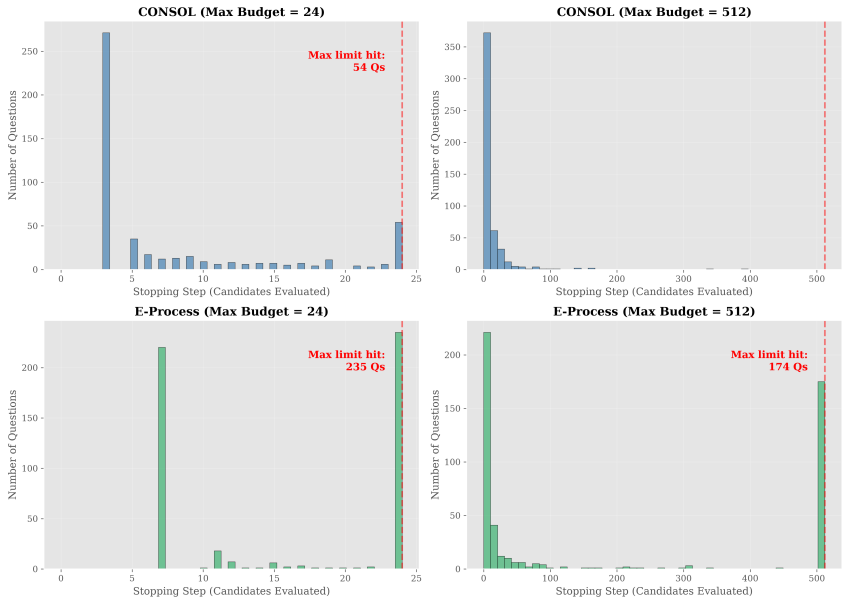


Figure 4: Distribution of stopping times on MATH500 ($N_{\max} = 512$). Exactly 174 out of 500 queries exhaust the evaluation budget, representing the subset of highly uncertain questions.

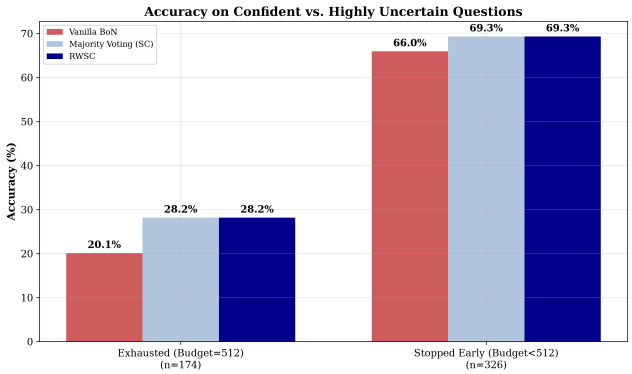


Figure 5: Performance comparison on confident (early stopping) versus highly uncertain (budget exhausted) subsets. While Vanilla BoN collapses on uncertain queries, RWSC and Borda safely fall back on the robustness of majority voting.