COUPLING SAMPLING WIDTH AND REASONING DEPTH FOR EFFICIENT LLM REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Inference-time scaling improves LLM reasoning, but its accuracy gains come with a rapidly growing token budget. Existing adaptive methods optimize a single axis: width-based approaches stop sampling once answers agree, risking premature commitment to an agreed-upon hallucination, while depth-based approaches prune trajectories in isolation, truncating hard but recoverable reasoning. Treating the two axes as independent leaves the budget-quality trade-off unresolved. We propose Dual-Dimensional Consistency (DDC), which couples both axes through one shared path-quality signal computed during generation. On the width axis, a confidence-weighted Bayesian rule terminates sampling only when the leading answer holds a high-confidence absolute majority; on the depth axis, trend-aware stratified pruning distinguishes persistent degradation from transient uncertainty via a spectral analysis of the confidence trajectory, gated per query without tuned constants. The budget freed by early termination is thereby redirected to the trajectories still worth extending. Across five reasoning benchmarks and five LLMs from 1.7B to 32B, DDC cuts token consumption by over $10\times$ while matching or exceeding strong adaptive baselines.

1 INTRODUCTION

Large Language Models (LLMs) have achieved remarkable performance across a wide range of tasks due to extensive pre-training and post-training (Devlin et al., 2019; Radford et al., 2021). However, these training-centric approaches incur substantial computational costs (Raffel et al., 2023; Stiennon et al., 2022). To address this, inference-time scaling (Wei et al., 2023; Wang et al., 2023; Wu et al., 2025; Xu et al., 2025) has emerged as an efficient alternative, significantly saving computational costs while maintaining and even enhancing performance (Deng et al., 2025).

To implement inference-time scaling, traditional static strategies like Self-Consistency (Wang et al., 2023), as illustrated in Figure 1(a), allocate a fixed budget like $B = 512$ to sample reasoning paths for majority voting to reach a consensus. This rigid paradigm is inefficient, as simple queries consume the same excessive resources as complex ones (Sriramanan et al., 2024). Fundamentally, this approach attempts to compensate for low path quality through massive samplings, inefficiently prioritizing quantity over the correctness of reasoning.

To address these limitations, recent studies explore adaptive inference strategies. Based on the classification in Figure 1(b), these methods optimize inference via two dimensions: inter-path width and intra-path depth. Width-based approaches manage sampling budget by terminating the entire sampling process upon early consensus. However, as models often exhibit hallucinations, assuming consistency equals correctness risks unsafe termination on consensus where paths agree on incorrect answers. Depth-based methods focus on quality of each path (Sharma & Chopra, 2025), pruning invalid paths via step-wise verification like token confidence (Snell et al., 2024; Huang et al., 2024). However, current pruning strategy often leads to sampling redundancy for simple queries and incorrect pruning of valid complex paths (Li et al., 2025a). Crucially, prevailing works treat width and depth as orthogonal objectives (Aggarwal et al., 2023), neglecting the intrinsic trade-off between sampling budget, generating excessive paths like Figure 1(b) and path quality, which leads to suboptimal resource allocation.

To bridge this gap, we introduce Dual-Dimensional Consistency(DDC), a unified framework designed to scale adaptive inference by jointly optimizing path quality and sampling budget. To minimize sam-

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

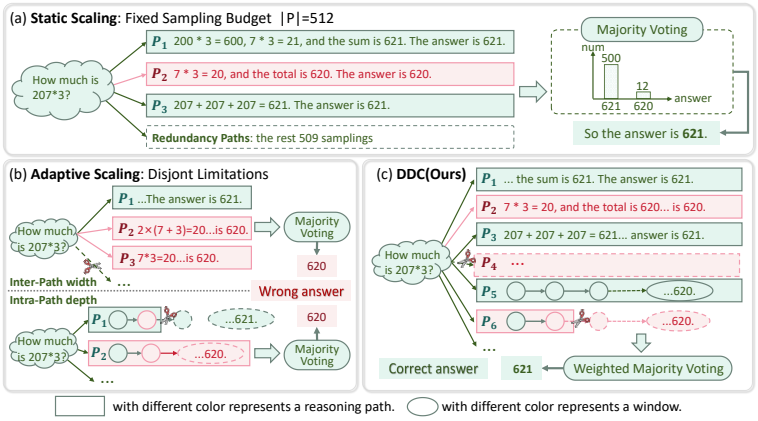


Figure 1: Comparisons of different methods. (a) is traditional static scaling method with fixed sampling budget of 512. (b) is the current adaptive scaling method of inter-path sampling width and intra-path generation depth. (c) is our method: a joint framework.

pling budget, we propose a Confidence-Weighted Bayesian Termination protocol. This mechanism integrates path-level quality into the consensus and permits early termination when the confidence-weighted support satisfies the posterior stopping criterion. Simultaneously, to refine path quality, we introduce Trend-Aware Stratified Pruning. Modeling each reasoning process as a dynamic signal to disentangle stable trends from stochastic noise (Li et al., 2025b; Yan et al., 2025), this pruning strategy aims to discard unreliable paths while retaining complex paths with recovering confidence. By unifying these components, DDC achieves balance between width of sampling and depth of individual paths, dynamically allocating resources to the most promising reasoning chains.

Evaluations across five challenging reasoning benchmarks illustrate the accuracy–cost trade-off. On AIME25 using Qwen3-4B, the reported DDC accuracy exceeds Self-Consistency by 15.6 percentage points at approximately $27\times$ fewer tokens. The arithmetic mean of the SC-to-DDC token ratios across the 25 model–task pairs is $13.7\times$; the ratio of their summed token counts is $8.5\times$.

The main contributions are highlighted as follows: **(1) Unified Optimization Framework.** We propose a framework that couples inter-path width with intra-path depth to maximize efficiency by jointly calibrating termination and pruning based on shared quality signals. **(2) Confidence-Weighted Bayesian Termination.** We formulate early stopping as a Bayesian sequential decision problem that weights sampling evidence with path quality to reduce premature termination on weakly supported consensus. **(3) Trend-Aware Stratified Pruning.** We develop a dynamic pruning strategy that distinguishes reasoning trends from noise, allowing the model to filter hallucinations while preserving valid reasoning paths. **(4) State-of-the-Art Performance.** Extensive evaluations show that DDC redefines the efficiency-accuracy trade-off, achieving a mean token-reduction factor above $10\times$ relative to SC across model–task pairs, with competitive average accuracy.

2 Related Works

Inference-Time Scaling. Recent advances in scaling have shifted from parameter scaling to inference-time scaling, revealing an inference-time scaling law (Snell et al., 2024; Brown et al., 2024; Wu et al., 2025) where increasing computational resources during inference consistently enhances performance (Hao et al., 2023). These methods are primarily categorized into training-based and training-free approaches. Training-based methods leverage reinforcement learning with verifiable rewards (RLVR) to induce long form reasoning patterns, as demonstrated in models such as DeepSeek-R1 (DeepSeek-AI et al., 2026) and other reasoning-optimized frameworks (Ye et al., 2025; Zuo et al., 2025). Training-free methods optimize reasoning during inference, typically through searching-based or sampling-based strategies. Searching-based methods (Yao et al., 2023; Hao et al., 2023; Xu et al., 2025), refine answers by exploring multiple paths, while sampling-based methods (Wang et al., 2023; Kang et al., 2025) generate and select from multiple candidate solutions. These methods improve

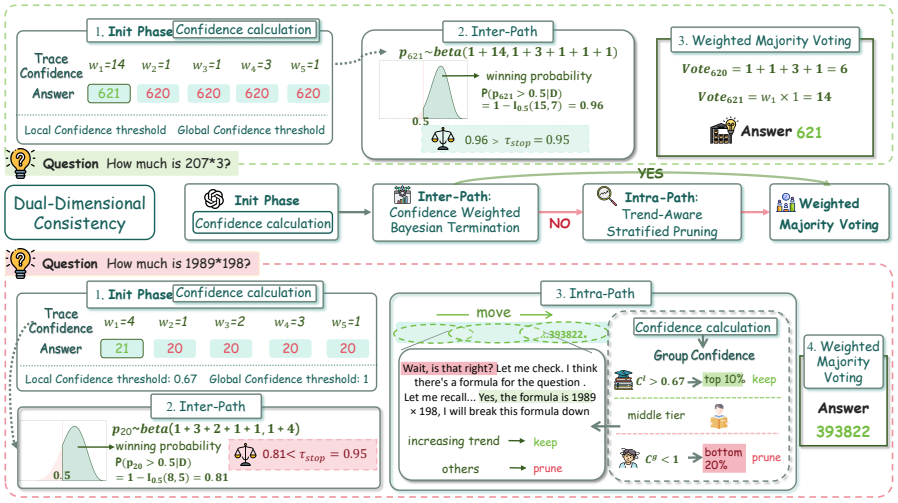


Figure 2: Illustration of the DDC inference process. Init Phase Confidence Calculation computes initial metrics to guide resource allocation. Inter-Path Bayesian Termination updates the Beta distribution using global path confidence as weights, triggering early stopping upon consensus. If generation continues, Intra-Path Trend-Aware Pruning monitors reasoning dynamics via a sliding window to dynamically discard invalid paths.

reasoning without retraining, though they often involve high computational costs (Madaan et al., 2023; Lan et al., 2024; Li et al., 2025b).

Adaptive Inference and Verification. Recent works try to dynamically adjust compute resources based on query difficulty. Inter-Path methods adaptively sample and aggregate paths (Li et al., 2024), using techniques like early stopping and difficulty aware policies (Wang et al., 2025; Kim et al., 2026) to reduce redundant samplings. Intra-Path methods optimize the depth of reasoning by guiding or pruning searches (Aggarwal et al., 2023) and depth control strategies MUR (Yan et al., 2025)) that avoid overthinking and hallucinations by monitoring uncertainty. These strategies treat inter-path width and intra-path depth as orthogonal objectives, neglecting the connection between quality and global budget.

Reasoning path Analysis. Monitoring the generation process is essential for reliable scaling. While token-level uncertainty can signal confidence (Quevedo et al., 2024; Zhang et al., 2025), raw scores are often too noisy for precise pruning (Yan et al., 2025). Drawing inspiration from signal processing (Li et al., 2025b), specifically Singular Spectrum Analysis, we denoise these paths. This allows us to rescue hard but correct paths that exhibit temporary confidence dips, a nuance overlooked by current static pruning and selection methods.

3 Problem Definition

LLM reasoning is a generation task: given query x, the model samples path y, tokens sequence $\{y_t\}_{t=1}^{T}$ according to the distribution $P_\theta(y \mid x)$, leading to an answer $u \in \mathcal{U}$. Let u_{gt} be the ground truth.

Static Scaling identifies the leader $u^* = \arg\max_{u \in \mathcal{U}} \Phi(u; S)$ from n independent paths $\mathcal{S} = \{y^{(p)}\}_{p=1}^{n}$, where $\Phi(u; \mathcal{S})$ is the consensus measuring the support for answer u within $\mathcal{S}$.

Adaptive Scaling defines this as a dual decision: (1) Inter-Path Termination in Width: Stop sampling when the posterior support for an absolute majority exceeds τ_{stop}. (2) Intra-Path Pruning in Depth: Prune $y^{(p)}$ at step t when its confidence signals indicate low quality. The objective is to minimize $\sum_{y \in \mathcal{A}_{\text{dyn}}} |y|$ while targeting $P(\Psi(S_{\text{dyn}}) = u_{\text{gt}}) \geq \delta_{\text{target}}$, where $\mathcal{A}_{\text{dyn}}$ includes all generated attempts, including pruned prefixes, and S_{dyn} contains the retained answer-bearing paths, where $\Psi(S_{\text{dyn}}) = u^*$ represents the final consensus answer, and δ_{target} is the reference scaling method's accuracy. This accuracy target is an empirical objective, not a guarantee implied by the stopping posterior.

4 METHODOLOGY

To balance sampling budget and path quality, we propose DDC, an adaptive framework utilizing a multi-granular Confidence Metrics.

As illustrated in Figure 2, DDC proceeds as follows: Init Phase Confidence Calculation first deploys a budget B_{init} to profile the initial confidence. Inter-Path Bayesian Termination then assesses, executing an early exit if the leading answer exhibits statistical sufficiency. If uncertainty remains, Intra-Path Trend-Aware Pruning continues generation up to a total budget of B attempts, including initialization and pruned attempts, while monitoring reasoning with a sliding window. Finally, Consensus derives the answer through confidence-weighted voting on retained answer-bearing paths; these paths are not externally verified.

4.1 MULTI-GRANULAR CONFIDENCE CALCULATION METRICS

To quantify reasoning quality, we adopt the confidence measurements proposed by (Fu et al., 2025), establishing a multi-granular confidence metrics that dynamically captures step-wise dynamics from token-level to path-level.

As illustrated in Figure 3, our metrics is defined as follows:

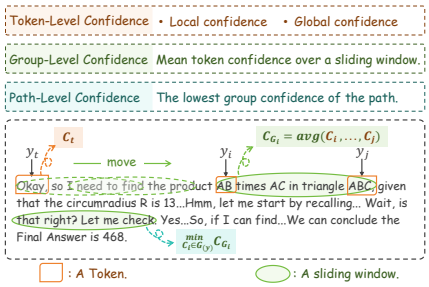


Figure 3: Illustrating the multi-granular confidence.

Token-Level Confidence. We measure the model's certainty at generation step t by two metrics: **(1)Local Confidence** C_t^l is defined as the maximum token probability: $C_t^l = P_t(1)$. **(2)Global Confidence** C_t^g captures richer distributional information from the top-k tokens by computing their negative average log-probability:

$$C_t^g = -\frac{1}{k}\sum_{j=1}^{k}\log P_t(j), \tag{1}$$

where $P_t(j)$ is the j-th largest token probability. This nonnegative score is used as a confidence proxy, not a calibrated correctness probability; it is not restricted to $[0, 1]$.

Group-Level Confidence. For a sliding window of tokens G_i starting at step i, the group confidence C_{G_i} is defined as the mean token confidence:

$$C_{G_i} = \frac{1}{|G_i|}\sum_{t=i}^{i+|G_i|-1} C_t, \tag{2}$$

applied separately to the local and global confidence channels.

Path-Level Confidence. The overall quality of a path y is its lowest group confidence:

$$C_{\text{path}}(y) = \min_{G_i \in \mathcal{G}(y)} C_{G_i}, \tag{3}$$

where $\mathcal{G}(y)$ is the set of sliding windows in y; for a path shorter than L, the entire path forms one window. We use the global channel $C_{G_i}^g$ for C_{path}, so the weight reflects the lowest-scoring segment and remains an unnormalized confidence proxy.

Init Phase. We generate a set of B_{init} paths. For each path, we compute the aforementioned confidence metrics to establish a query-specific quality baseline, forming the initial evidence.

4.2 INTER-PATH: CONFIDENCE-WEIGHTED BAYESIAN TERMINATION

To minimize sampling redundancy without the overhead of high-dimensional Dirichlet integration (Aggarwal et al., 2023), we reframe consensus as a binary competition between the leading candidate u^* and the aggregate alternatives, simplifying the model to a Beta distribution, enabling efficient real-time termination.

Bayesian Sequential Decision Formulation. We frame the adaptive scaling of sampling width as a generalized Bayesian sequential decision problem. After initialization, $p_u \mid \mathcal{D} \sim \text{Beta}(\alpha_u, \beta_u)$ models the confidence-weighted support share for candidate u against all alternatives. The parameters accumulate supporting and opposing evidence; this posterior quantifies consensus uncertainty, not the probability that u is correct.

Adaptive Stopping Rule. To determine the termination point, we utilize the cumulative distribution function (CDF) of $\text{Beta}(\alpha_u, \beta_u)$:

$$I_\gamma(\alpha_u, \beta_u) = \frac{\int_0^\gamma t^{\alpha_u - 1}(1-t)^{\beta_u - 1} dt}{\int_0^1 t^{\alpha_u - 1}(1-t)^{\beta_u - 1} dt}, \tag{4}$$

which gives the posterior probability that $p_u \leq \gamma$. Generation terminates when the complementary probability for the leading answer $u^* = \arg\max_u \alpha_u$ exceeds the stopping threshold:

$$\mathbb{P}(p_{u^*} > \gamma \mid \mathcal{D}) = 1 - I_\gamma(\alpha_{u^*}, \beta_{u^*}) > \tau_{\text{stop}}. \tag{5}$$

We set $\gamma = 0.5$ to test absolute-majority support rather than a lead over only the runner-up, and $\tau_{\text{stop}} = 0.95$ as the posterior stopping threshold. This is a consensus criterion, not a 95% guarantee of answer correctness.

Confidence-Weighted Parameter Update. To further enhance the reliability of this early exit, we introduce a confidence-weighted Bayesian update mechanism. Instead of treating each sample equally, we assign a weight $w_i = C_{\text{path}}(y^{(i)})$ to each reasoning path. This integrates the path-quality proxy directly into the posterior update. It can reduce the influence of low-confidence votes, but systematic high-confidence errors can still trigger an incorrect early exit.

For a candidate answer u, we define the accumulated evidence α_u in support of u and counter-evidence β_u in support of all competitors as follows:

$$\alpha_u = 1 + \sum_i \mathbb{I}\{\text{ans}_i = u\} \cdot w_i, \tag{6}$$

$$\beta_u = 1 + \sum_j \mathbb{I}\{\text{ans}_j \neq u\} \cdot w_j, \tag{7}$$

where the constant 1 represents a Laplace prior, corresponding to a uniform prior $\text{Beta}(1, 1)$ to account for initial uncertainty. This provides a symmetric initial state before any evidence is collected; subsequent concentration depends on the accumulated weight scale. Detailed analysis theory is in Appendix B.

4.3 Intra-Path: Trend-Aware Stratified Pruning

To optimize path quality for adaptive inference scaling, it is insufficient to simply increase the number of samples, as indiscriminate scaling introduces excessive noise and computational waste. We propose Trend-Aware Stratified Pruning to preferentially retain promising reasoning paths and prune paths with low confidence or anomalous trends before the final vote. This hierarchical gating mechanism transitions from static scalar thresholds to a dynamical analysis of the reasoning process, allowing the model to adaptively allocate compute based on the evolving trend of each reasoning path. We categorize paths into three tiers based on the trade-off between computational cost and information exploitation.

First, we calculate the group confidence C_{G_m} using local confidence. Paths maintaining confidence higher than top-10% of the init distribution bypass subsequent analysis and proceed with reasoning on the insight that high local confidence correlates with correct reasoning (Fu et al., 2025). Then we calculate C_{G_m} utilizing global confidence with more comprehensive information and prune paths below the initialization distribution's 20th-percentile threshold. The critical challenge lies in the middle tier paths, which contain a mixture of valid difficult reasoning and hallucinations. Static averaging fails here, as it cannot distinguish between a temporary confidence dip during hard problem-solving and the chaotic, stagnant confidence characteristic of a hallucination.

To resolve this, we map the reasoning process into a dynamical phase space. We construct a position-velocity state vector $\mathbf{z}_t \in \mathbb{R}^2$ for each step t within a window of size L:

$$\mathbf{z}_t = \begin{bmatrix} \dfrac{C_t^g - \mu_c}{\sigma_c} \\ \dfrac{C_t^g - C_{t-1}^g}{\sigma_c} \end{bmatrix}, \tag{8}$$

where C_t^g is the global confidence, μ_c is the window mean, and σ_c is its standard deviation. For nondegenerate windows, normalization makes this representation invariant to positive affine rescaling, not arbitrary monotone transformations. The components represent relative position and velocity. We form $\mathbf{Z} \in \mathbb{R}^{L\times 2}$ and decompose its second-moment matrix $\mathbf{\Sigma} = \frac{1}{L-1}\mathbf{Z}^\top\mathbf{Z}$ into eigenvalues $\lambda_1 \geq \lambda_2$ and principal eigenvector $\vec{v_1}$. A large eigenvalue gap indicates directional structure, whereas similar eigenvalues indicate weaker directionality; neither pattern alone identifies correctness or hallucination. We define the **Structural Instability Score** $\mathcal{R}$ to penalize weak directionality and negative momentum:

$$\mathcal{R} = \left(1 - \frac{\lambda_1 - \lambda_2}{\lambda_1 + \lambda_2}\right) + \eta \cdot \mathbb{I}(\text{align} < 0) \cdot \text{align}^2, \tag{9}$$

where $\text{align} = |v_{1,x}| \cdot \text{sign}(\bar{v})$, $|v_{1,x}|$ represents the alignment of the path's primary trend with the confidence axis, and $\bar{v}$ is the mean normalized velocity. The align metric captures the principal-axis orientation and the velocity sign. We set the first within-window difference to zero; when $\sigma_c = 0$ or $\lambda_1 + \lambda_2 = 0$, we define the trend score as $\mathcal{R} = 0$, leaving confidence gating to handle the constant window. Crucially, we employ the mean velocity $\bar{v}$ as the directional pivot for align because the rate of change in confidence serves as a leading indicator of reasoning failure rather than position.

If the dominant trend is a sustained decrease in confidence where $\bar{v} < 0$, we impose a quadratic penalty. This increases the risk score of declining paths relative to recovering paths with otherwise similar structure; pruning still depends on the gates and the query-specific risk threshold.

Finally, we establish the pruning criterion based on $\mathcal{R}$. While static percentile thresholds suffice for filtering clear-cut confidence extremes as in our initial gating, they are ill-suited for this middle tier where the baseline instability varies significantly across queries. Imposing a fixed quota here risks indiscriminately pruning valid paths in complex tasks or retaining noise in trivial ones. To ensure robust adaptivity, we formulate pruning as a query-specific outlier detection problem. We derive the dynamic threshold τ_{risk} using Tukey Fences (Tukey, 1977), which defines a statistical boundary based on the dispersion of the init population:

$$\tau_{\text{risk}} = Q_3 + 1.5 \cdot \text{IQR}, \tag{10}$$

where Q_3 denotes the third quartile and $\text{IQR} = Q_3 - Q_1$. By calibrating the threshold against the population's interquartile range, this method automatically scales with the structural variance of the reasoning process, effectively isolating anomalous paths $\mathcal{R} > \tau_{\text{risk}}$ while accommodating the intrinsic difficulty of the task.

4.4 Final Consensus

The final answer is determined by applying weighted voting on the multiset of retained answer-bearing paths $\mathcal{S} = \mathcal{Y}_{init} \cup \mathcal{Y}_{valid}$, preserving repeated samples.

$$\hat{y} = \arg\max_u \sum_{y^{(i)} \in \mathcal{S}} \mathbb{I}\{\text{ans}_i = u\} \cdot C_{\text{path}}(y^{(i)}), \tag{11}$$

where each answer's support depends on both the number of supporting paths and their confidence weights. The weighted argmax need not constitute an absolute majority when the budget is exhausted.

5 Experiments

5.1 Experimental Setup

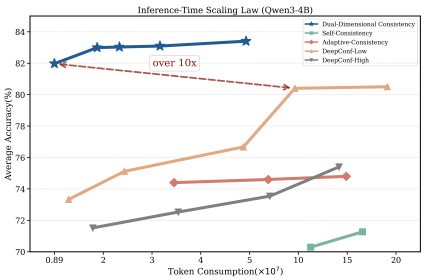


Figure 4: Inference-time scaling law on Qwen3-4B-Instruct. The horizontal axis denotes the token consumption, while the vertical axis represents the average accuracy on 5 benchmarks.

Datasets and Models. We evaluate on five challenging reasoning benchmarks: MATH-500 (Hendrycks et al., 2021), AMC23, AIME24, AIME25, and GPQA-diamond (Rein et al., 2023). To make fair comparison, we evaluate on the same LLMs following previous works (Fu et al., 2025; Yan et al., 2025), specifically on Qwen 3 series and DeepSeek-R1-0528-Qwen3-8B. Detailed descriptions are provided in Appendix A.1.

Baselines. We compare our method with four established Self-Consistency approaches: Self-Consistency (SC) (Wang et al., 2023), which samples multiple reasoning paths and selects the final answer via majority voting; Adaptive-Consistency (AC) (Aggarwal et al., 2023), which dynamically adjusts the sample count based on answer consensus; and confidence-based strategies, DeepConf-Low (D-L) and DeepConf-High (D-H) (Fu et al., 2025), which employ different confidence thresholds to balance exploration and exploitation.

Metrics. We report answer accuracy under a maximum sampling budget of 512 and token consumption in 10^7 units. DDC's final aggregator is the weighted vote in Section 4.4; $B_{\text{init}} = 16$ denotes the paths generated per query at initialization, not independent repetitions.

Implementation Details. We set $T = 0.6$, top-$p = 0.95$, and sampling budget $B = 512$ unless otherwise noted. The init size is $B_{\text{init}} = 16$ and $L(2048)$ denoting the sliding window size for group confidence is 2048 tokens. Further implementation details are in Appendix A.3.

5.2 Main Results

DDC strikes a superior balance between effectiveness and efficiency over strong baselines. As shown in Table 1, DDC offers a favorable reported trade-off across diverse benchmarks and model scales, from Qwen3-1.7B to 32B and the DeepSeek distillation. On Qwen3-4B, its reported average accuracy exceeds Self-Consistency by 8.6 percentage points and DeepConf-Low by 2.9 points. On AIME25, the reported gain over Self-Consistency is 15.6 points. DDC does not lead in accuracy in every model–task pair.

DDC significantly minimizes computational overhead by reducing token consumption. Using the ratio of summed token counts across the 25 model–task pairs, DDC reduces the reported token total by $8.5\times$ relative to Self-Consistency and $3.1\times$ relative to Adaptive-Consistency. On AIME25 with Qwen3-4B, the SC-to-DDC token ratio is approximately $27\times$. Token savings and latency savings are distinct: for example, on Qwen3-32B/AIME25, the measured generation-time ratio relative to DeepConf-Low is $12.4\times$. Detailed complexity and timing analyses are provided in Appendix C.

5.3 On the Inference-Time scaling

Figure 4 presents the accuracy–token scaling curves on Qwen3-4B. These curves illustrate the reported trade-off across token-consumption levels rather than a guarantee of superiority at every budget. Detailed results are in Appendix C.

5.4 Superiority over compute-intensive methods.

Table 2 demonstrates that DDC surpasses computationally exhaustive strategies like Predictive Decoding (Ma et al., 2024) and ϕ-Decoding (Xu et al., 2025). On AIME25 with Qwen3-4B, its reported accuracy exceeds ϕ-Decoding by 15.0 percentage points at $29\times$ fewer tokens. On GPQA-diamond with DeepSeek-R1-0528-Qwen3-8B, ϕ-Decoding is 3.0 points more accurate but uses $6.9\times$ as many tokens, illustrating a task-dependent trade-off.

5.5 Ablation Studies

Table 3 reveals that DDC's performance stems from a functional interdependency between Inter-path and Intra-path.

Table 1: Comprehensive evaluation across reasoning benchmarks. Acc. is reported answer accuracy (%); Tokens is the recorded token consumption ($\times 10^7$) under a maximum sampling budget of 512. Best reported results are in **bold**.

Method	MATH500		AMC23		AIME24		AIME25		GPQA-d		Avg.	
	Acc.↑	Tokens↓	Acc.↑	Tokens↓	Acc.↑	Tokens↓	Acc.↑	Tokens↓	Acc.↑	Tokens↓	Acc.↑	Tokens↓
Qwen3-1.7B												
SC	76.5	44.9	94.5	18.7	74.1	49.0	49.5	47.5	40.2	186.4	67.0	69.3
AC	76.7	24.8	94.3	10.2	74.0	19.3	49.4	17.8	40.1	69.7	66.9	28.4
D-L	84.0	26.2	92.5	9.7	**76.7**	9.6	**53.3**	9.3	40.4	126.6	69.4	36.3
D-H	83.8	29.6	91.1	15.7	**76.7**	16.8	53.0	23.4	40.3	166.5	69.0	50.4
DDC	**89.4**	**10.2**	**95.0**	**1.5**	73.3	**3.8**	50.0	**3.7**	**42.9**	**20.6**	**70.1**	**8.0**
Qwen3-4B												
SC	83.0	48.6	96.0	16.6	80.0	38.4	66.5	41.0	48.5	80.4	74.8	45.0
AC	83.2	18.1	96.0	4.5	79.9	12.8	66.4	15.6	48.5	23.7	74.8	14.9
D-L	87.0	14.1	**100.0**	0.8	83.1	14.1	73.3	18.3	59.0	48.2	80.5	19.1
D-H	87.3	30.6	96.2	10.4	80.2	24.2	66.7	25.8	48.7	50.6	75.8	28.3
DDC	**92.8**	**7.5**	**100.0**	**0.5**	**83.3**	**1.2**	**82.1**	**1.5**	**59.1**	**13.7**	**83.4**	**4.9**
Qwen3-8B												
SC	85.5	51.2	96.5	15.4	80.0	23.2	78.1	27.7	63.7	74.7	80.8	38.4
AC	85.6	13.0	96.5	6.1	80.0	12.2	70.0	9.2	63.3	24.8	79.1	13.1
D-L	87.9	11.4	96.7	12.5	**80.4**	9.0	76.7	13.1	**65.2**	33.1	81.3	15.8
D-H	87.8	19.3	96.0	18.3	80.1	13.3	**78.5**	19.9	63.8	49.4	81.3	24.0
DDC	**92.8**	**7.3**	**97.5**	**1.0**	80.0	**1.3**	73.3	**2.2**	63.6	**13.7**	**81.4**	**5.1**
Qwen3-32B												
SC	90.0	76.8	**97.5**	10.2	84.8	20.0	80.1	24.3	71.1	74.4	84.7	41.1
AC	90.1	21.5	97.0	3.4	84.9	10.6	80.0	9.1	70.1	23.7	84.4	13.7
D-L	92.7	35.8	97.3	7.9	89.4	7.8	80.2	11.4	70.9	32.1	86.1	19.0
D-H	91.8	40.0	96.8	9.3	86.4	8.8	**80.2**	16.1	70.0	41.6	85.0	23.2
DDC	**93.6**	**6.7**	**97.5**	**0.6**	**93.3**	**1.3**	79.7	**1.8**	**72.2**	**14.1**	**87.2**	**4.9**
DeepSeek-R1-0528-Qwen3-8B												
SC	91.1	53.8	93.5	14.1	86.7	35.5	82.3	40.1	**72.5**	99.2	85.2	48.5
AC	91.0	17.9	93.5	4.7	86.6	11.8	82.3	13.3	72.4	38.0	85.2	17.1
D-L	91.3	16.4	93.9	11.0	**92.1**	7.8	**86.4**	12.4	71.7	34.6	**87.1**	16.4
D-H	91.5	25.6	94.0	9.7	86.7	14.5	81.4	23.7	72.4	69.1	85.2	28.5
DDC	**93.8**	**9.5**	**100.0**	**0.8**	86.7	**1.9**	83.3	**2.4**	69.6	**13.9**	86.7	**5.7**

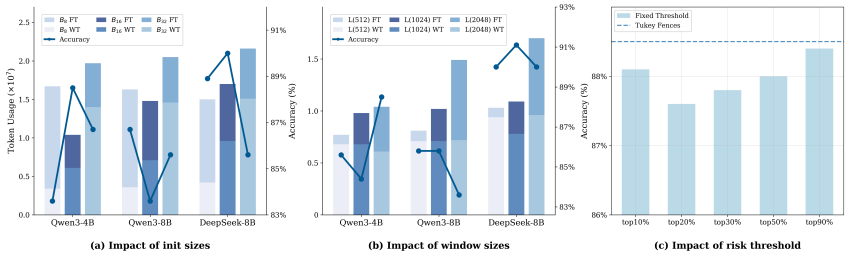
Table 2: Comparison with compute-intensive methods.

Method	GPQA-diamond		AIME25	
	Acc.	Tokens	Acc.	Tokens
Qwen3-4B				
DDC	**59.1**	**13.7**	**82.1**	**1.5**
Predictive Decoding + SC	54.8	94.3	66.8	48.9
ϕ-Decoding + SC	56.6	86.9	67.1	43.3
DeepSeek-R1-0528-Qwen3-8B				
DDC	69.6	**13.9**	**83.3**	**2.4**
Predictive Decoding + SC	69.5	98.6	78.0	57.1
ϕ-Decoding + SC	**72.6**	95.8	79.2	52.6

Table 3: Ablation Studies on Qwen3-4B and DeepSeek-R1-0528-Qwen3-8B.

Method	MATH500		AIME25	
	Acc.	Tokens	Acc.	Tokens
Qwen3-4B				
DDC	**92.8**	**7.5**	**82.1**	**1.5**
w/o Weighted Bayesian	90.2	25.6	80.0	6.1
w/o Inter-path Scaling	88.1	29.7	77.6	21.6
w/o Trend-Aware	92.0	7.7	75.3	1.1
w/o Intra-path Pruning	85.5	18.4	70.9	10.3
DeepSeek-R1-0528-Qwen3-8B				
DDC	**93.8**	**9.5**	**83.3**	**2.4**
w/o Weighted Bayesian	91.9	10.1	82.6	6.6
w/o Inter-path Scaling	92.3	20.9	81.6	19.3
w/o Trend-Aware	93.6	9.0	83.1	2.0
w/o Intra-path Pruning	91.2	18.4	82.5	14.9

Weighted Bayesian accelerates convergence. On Qwen3-4B/MATH500, frequency-based termination (*w/o Weighted Bayesian*) uses approximately $3.4\times$ as many tokens as full DDC (25.6 versus



(a) Impact of init sizes (b) Impact of window sizes (c) Impact of risk threshold

Figure 5: Hyperparameter sensitivity analysis. We report average accuracy and token consumption on competition benchmarks: (a) Impact of init sizes. (b) Impact of window sizes. (c) Comparison between adaptive Tukey Fences and static percentage-based strategies.

7.5), with a 2.6-point lower accuracy (90.2% versus 92.8%). This comparison supports confidence weighting as a useful component of the shared quality signal.

Inter-path Scaling optimizes resource allocation. On Qwen3-4B/MATH500, disabling adaptive budgeting (*w/o Inter-path Scaling*) uses approximately $4.0\times$ as many tokens (29.7 versus 7.5), with a 4.7-point lower accuracy (88.1% versus 92.8%). This supports adaptive stopping in the reported setting, without implying that early commitment always preserves correctness.

Trend-aware analysis mitigates hallucinations. On Qwen3-4B/AIME25, replacing dynamic signal analysis with static confidence pruning (*w/o Trend-Aware*) lowers reported accuracy by 6.8 percentage points (82.1% to 75.3%). This supports the usefulness of trend information, but does not establish perfect separation of valid reasoning and hallucinations.

Intra-path Pruning prevents error propagation. On Qwen3-4B/MATH500, removing intra-path pruning lowers reported accuracy from 92.8% to 85.5%, a 7.3-point drop, the largest among these ablations. This supports pruning as a useful component rather than proving that every pruned path is incorrect.

Hyperparameter Analysis. We analyze init sizes B_{init} and window granularity L across three competition datasets in Figure 5. The settings $B_{\text{init}} = 16$, $L = 2048$ provide the default trade-off, although the best setting varies across models and datasets; for the models that underperform in Table 1, such as Qwen3-8B and DeepSeek-R1-0528-Qwen3-8B, adjusting B_{init} and L further improves accuracy. Our adaptive τ_{risk} thresholding consistently outperforms static percentage-based strategies, and since it is derived per query via Tukey Fences rather than tuned manually, the main comparison uses one shared hyperparameter set from 1.7B to 32B; this does not guarantee optimal settings for an unseen model or dataset. Detailed experiments are provided in Appendix C.

6 Conclusion

We address the inefficiency of static and disjoint adaptive inference scaling with DDC, a unified framework that balances budget and quality. By coupling a Confidence-Weighted Bayesian termination protocol with Trend-Aware Stratified Pruning, DDC concentrates computation on high quality reasoning chains while identifying and discarding hallucinations, achieving state-of-the-art performance and offering a scalable pathway for deploying reliable reasoning in resource-constrained environments.

AI Use Statement

Generative AI tools assisted language editing, reference checks, and consistency review of notation, theoretical arguments, and reported calculations. These checks did not produce new experimental measurements. The authors are responsible for reviewing and verifying the final text, theoretical claims, experimental records, and artifacts.

Ethics Statement

This work advances efficient inference for large language models by jointly optimizing sampling budget and reasoning-path quality. It uses only publicly available models and the public benchmarks described in Section 5.1, and involves no human subjects, private data, or new data collection. By cutting the token cost of reliable reasoning, DDC reduces the energy and monetary footprint of test-time scaling. As analyzed in Appendix B.3, DDC relies on the correlation between confidence and correctness and may terminate early on a confidently wrong consensus; high-stakes deployments should therefore pair it with external verification. To the best of our knowledge, this work poses no additional ethical or societal risks.

Reproducibility Statement

The main comparison uses publicly available models and the five public benchmarks in Section 5.1, with default hyperparameters in Appendix A.3. Sensitivity analyses and the HumanEval proof of concept use the explicitly stated alternative settings. The confidence metrics, termination rule, and consensus are defined in Section 4; the complete procedure is given as Algorithms 1–4 in Appendix D; and all assumptions and proofs appear in Appendix B. Dataset descriptions, metric definitions, latency protocols, and sensitivity studies are provided in Appendices A and C.

References

Pranjal Aggarwal, Aman Madaan, Yiming Yang, and Mausam. Let's sample step by step: Adaptive-consistency for efficient reasoning and coding with llms, 2023. URL https://arxiv.org/abs/2305.11860.

José M. Bernardo and Adrian F. M. Smith. *Bayesian Theory*. Wiley Series in Probability and Statistics. John Wiley & Sons, 1994.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling, 2024. URL https://arxiv.org/abs/2407.21787.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code, 2021. URL https://arxiv.org/abs/2107.03374.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia

He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2026. URL https://arxiv.org/abs/2501.12948.

Zeshuai Deng, Guohao Chen, Shuaicheng Niu, Hui Luo, Shuhai Zhang, Yifan Yang, Renjie Chen, Wei Luo, and Mingkui Tan. Test-time model adaptation for quantized neural networks, 2025. URL https://arxiv.org/abs/2508.02180.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL https://arxiv.org/abs/1810.04805.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Yichao Fu, Xuewei Wang, Yuandong Tian, and Jiawei Zhao. Deep think with confidence, 2025. URL https://arxiv.org/abs/2508.15260.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pp. 1321–1330, 2017.

J. M. Hammersley and D. C. Handscomb. *Monte Carlo Methods*. Methuen & Co Ltd, London, 1964.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. Reasoning with language model is planning with world model, 2023. URL https://arxiv.org/abs/2305.14992.

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, 2021. URL https://arxiv.org/abs/2103.03874.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet, 2024. URL https://arxiv.org/abs/2310.01798.

Joseph G. Ibrahim and Ming-Hui Chen. Power prior distributions for regression models. *Statistical Science*, 15(1):46–60, 2000.

Joseph G. Ibrahim, Ming-Hui Chen, Yeongjin Gwon, and Fang Chen. The power prior: theory and applications. *Statistics in Medicine*, 34(28):3724–3749, 2015.

Zhewei Kang, Xuandong Zhao, and Dawn Song. Scalable best-of-n selection for large language models via self-certainty, 2025. URL https://arxiv.org/abs/2502.18581.

Junseok Kim, Nakyeong Yang, Kyungmin Min, and Kyomin Jung. Reliability-aware adaptive self-consistency for efficient sampling in llm reasoning, 2026. URL https://arxiv.org/abs/2601.02970.

Tian Lan, Wenwei Zhang, Chen Xu, Heyan Huang, Dahua Lin, Kai Chen, and Xian ling Mao. Criticeval: Evaluating large language model as critic, 2024. URL https://arxiv.org/abs/2402.13764.

Bo Li, Tian Tian, Zhenghua Xu, Hao Cheng, Shikun Zhang, and Wei Ye. Modeling uncertainty trends for timely retrieval in dynamic rag, 2025a. URL https://arxiv.org/abs/2511.09980.

Xianliang Li, Jun Luo, Zhiwei Zheng, Hanxiao Wang, Li Luo, Lingkun Wen, Linlong Wu, and Sheng Xu. On the performance analysis of momentum method: A frequency domain perspective, 2025b. URL https://arxiv.org/abs/2411.19671.

Yiwei Li, Peiwen Yuan, Shaoxiong Feng, Boyuan Pan, Xinglin Wang, Bin Sun, Heda Wang, and Kan Li. Escape sky-high cost: Early-stopping self-consistency for multi-step reasoning, 2024. URL https://arxiv.org/abs/2401.10480.

Chang Ma, Haiteng Zhao, Junlei Zhang, Junxian He, and Lingpeng Kong. Non-myopic generation of language models for reasoning and planning, 2024. URL https://arxiv.org/abs/2410.17195.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback, 2023. URL https://arxiv.org/abs/2303.17651.

OpenAI. gpt-oss-120b & gpt-oss-20b model card, 2025. URL https://arxiv.org/abs/2508.10925.

Ernesto Quevedo, Jorge Yero, Rachel Koerner, Pablo Rivas, and Tomas Cerny. Detecting hallucinations in large language model generation: A token probability approach, 2024. URL https://arxiv.org/abs/2405.19648.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2023. URL https://arxiv.org/abs/1910.10683.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. Gpqa: A graduate-level google-proof q&a benchmark, 2023. URL https://arxiv.org/abs/2311.12022.

Aman Sharma and Paras Chopra. Think just enough: Sequence-level entropy as a confidence signal for llm reasoning, 2025. URL https://arxiv.org/abs/2510.08146.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters, 2024. URL https://arxiv.org/abs/2408.03314.

Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. Llm-check: Investigating detection of hallucinations in large language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 34188–34216. Curran Associates, Inc., 2024. doi: 10.52202/079017-1077. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/3c1e1fdf305195cd620c118aaa9717ad-Paper-Conference.pdf.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback, 2022. URL https://arxiv.org/abs/2009.01325.

John W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, Reading, MA, 1977.

A. Wald. Sequential tests of statistical hypotheses. *Annals of Mathematical Statistics*, 16(2):117–186, 1945.

Xinglin Wang, Shaoxiong Feng, Yiwei Li, Peiwen Yuan, Yueqi Zhang, Chuyi Tan, Boyuan Pan, Yao Hu, and Kan Li. Make every penny count: Difficulty-adaptive self-consistency for cost-efficient reasoning, 2025. URL https://arxiv.org/abs/2408.13457.

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, 2023. URL https://arxiv.org/abs/2203.11171.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL https://arxiv.org/abs/2201.11903.

L. Weiss. Sequential estimation. *Wiley StatsRef: Statistics Reference Online*, 2014.

Thomas D. Wickens. Categorical data analysis. *Annual Review of Psychology*, 49(1):537–557, 1998.

Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. Inference scaling laws: An empirical analysis of compute-optimal inference for problem-solving with language models, 2025. URL https://arxiv.org/abs/2408.00724.

Fangzhi Xu, Hang Yan, Chang Ma, Haiteng Zhao, Jun Liu, Qika Lin, and Zhiyong Wu. ϕ-decoding: Adaptive foresight sampling for balanced inference-time exploration and exploitation, 2025. URL https://arxiv.org/abs/2503.13288.

Hang Yan, Fangzhi Xu, Rongman Xu, Yifei Li, Jian Zhang, Haoran Luo, Xiaobao Wu, Luu Anh Tuan, Haiteng Zhao, Qika Lin, and Jun Liu. Mur: Momentum uncertainty guided reasoning for large language models, 2025. URL https://arxiv.org/abs/2507.14958.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models, 2023. URL https://arxiv.org/abs/2305.10601.

Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. Limo: Less is more for reasoning, 2025. URL https://arxiv.org/abs/2502.03387.

Tunyu Zhang, Haizhou Shi, Yibin Wang, Hengyi Wang, Xiaoxiao He, Zhuowei Li, Haoxian Chen, Ligong Han, Kai Xu, Huan Zhang, Dimitris Metaxas, and Hao Wang. Tokur: Token-level uncertainty estimation for large language model reasoning, 2025. URL https://arxiv.org/abs/2505.11737.

Jiaru Zou, Ling Yang, Jingwen Gu, Jiahao Qiu, Ke Shen, Jingrui He, and Mengdi Wang. Reasonflux-prm: Trajectory-aware prms for long chain-of-thought reasoning in llms, 2025. URL https://arxiv.org/abs/2506.18896.

Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Yuchen Zhang, Xinwei Long, Ermo Hua, Biqing Qi, Youbang Sun, Zhiyuan Ma, Lifan Yuan, Ning Ding, and Bowen Zhou. Ttrl: Test-time reinforcement learning, 2025. URL https://arxiv.org/abs/2504.16084.

A Implementation Details

For all experiments, we set the temperature to 0.6 and top-p to 0.95 unless otherwise specified.

A.1 Datasets Description

We evaluate on five challenging reasoning benchmarks:

- **MATH-500**: A subset of 500 problems from the MATH dataset (Hendrycks et al., 2021) covering algebra, geometry, calculus, and number theory. Problems require multi-step reasoning and symbolic manipulation.
- **AMC23**: The 2023 American Mathematics Competition problems, comprising 40 problems from the benchmark's AMC 2023 collection.
- **AIME24**: The 2024 American Invitational Mathematics Examination (AIME) problems, comprising 30 advanced problems that require deep mathematical insight and creative problem-solving.
- **AIME25**: The 2025 AIME problems, comprising 30 problems across AIME I and II that test advanced mathematical reasoning.
- **GPQA-diamond**: A high-quality subset of the Graduate-Level Google-Proof Q&A (GPQA) dataset (Rein et al., 2023), containing 198 questions across physics, chemistry, and biology that require expert-level knowledge and reasoning.

All datasets are in English. For each question, the model generates step-by-step reasoning before producing the final answer.

A.2 Metrics Calculation

We report two main metrics:

Accuracy (Acc.@K) : Accuracy is the fraction of evaluated questions whose selected answer matches the ground truth, under a maximum of K generation attempts ($K = 512$ by default). DDC's final aggregator is the weighted vote in Section 4.4.

Token Consumption : The intended metric sums generated tokens over all attempts and questions in the evaluated dataset, including initialization and pruned prefixes. Counts are reported in units of 10^7.

Additional Metrics : For the inference-time scaling analysis in Figure 4, we compute the average accuracy across all five benchmarks at various token budget levels, ranging from 0.89×10^7 to 20×10^7 tokens.

A.3 Hyperparameter Settings

The key hyperparameters for Dual-Dimensional Consistency are:

- **init size** (B_{init}): 16 samples for all experiments unless otherwise specified.
- **Sampling budget** (B): 512 samples for all main experiments.
- **Sliding window size** (L): 2048 tokens for group confidence calculation.
- **Majority threshold** (γ): 0.5 denoting absolute majority.
- **Stopping confidence threshold** (τ_{stop}): 0.95.
- **Trend analysis window size**: 2048 tokens.
- **Penalty coefficient** (η): 0.5 in the Structural Instability Score $\mathcal{R}$.

B THEORETICAL ANALYSIS AND PROOFS

In this section, we provide a formal analysis of the Dual-Dimensional Consistency (DDC) framework, grounding our approach in established statistical literature.

B.1 INTER-PATH: GENERALIZED BAYESIAN TERMINATION

B.1.1 FORMULATION AS GENERALIZED POSTERIOR

To integrate finite nonnegative confidence weights $w_n = C_{\text{path}}(y^{(n)})$ into consensus, we use a power-likelihood update inspired by the Power Prior framework (Ibrahim & Chen, 2000). The raw scores are not restricted to $[0, 1]$; they define evidence weights rather than correctness probabilities.

Consider the Bernoulli likelihood for $u_n \in \{0, 1\}$, where $u_n = 1$ indicates support for a fixed candidate. For the efficiency analysis below, that candidate is taken to be the correct answer; the algorithm does not observe correctness. The weighted log-likelihood is:

$$\log L_w(\mathcal{D}_N \mid \theta) = \sum_{n=1}^{N} w_n \left[u_n \log \theta + (1 - u_n) \log(1 - \theta)\right]. \tag{12}$$

This is a generalized Bayesian update with confidence-weighted observations. Unlike importance sampling (Hammersley & Handscomb, 1964), these weights are not proposal-to-target density ratios. Combined with a conjugate Beta prior $\text{Beta}(\alpha_0, \beta_0)$, this yields a posterior:

$$\theta \mid \mathcal{D}_N \sim \text{Beta}\left(\alpha_0 + \sum_{n=1}^{N} w_n u_n, \beta_0 + \sum_{n=1}^{N} w_n (1 - u_n)\right). \tag{13}$$

Writing $A = \alpha_0 + \sum_n w_n u_n$ and $B' = \beta_0 + \sum_n w_n (1 - u_n)$, the posterior variance is $AB'/[(A + B')^2(A + B' + 1)]$. For a fixed interior posterior mean and growing total evidence, it scales as $O((\sum_n w_n)^{-1})$. Thus weight scale affects concentration as well as relative support; a universal safety guarantee does not follow from nonnegativity alone.

B.1.2 EFFICIENCY ANALYSIS VIA SAMPLE COMPLEXITY

Motivated by sequential analysis (Wald, 1945), we compare signal-to-noise sample-complexity proxies for Confidence-Weighted (CoW) and Frequency-Based (FrQ) evidence. Let $p = \Pr(u_n = 1)$ and $q = 1 - p$ for a fixed sampling distribution. This comparison does not establish an ordering of the actual Beta-CDF stopping times under adaptive pruning.

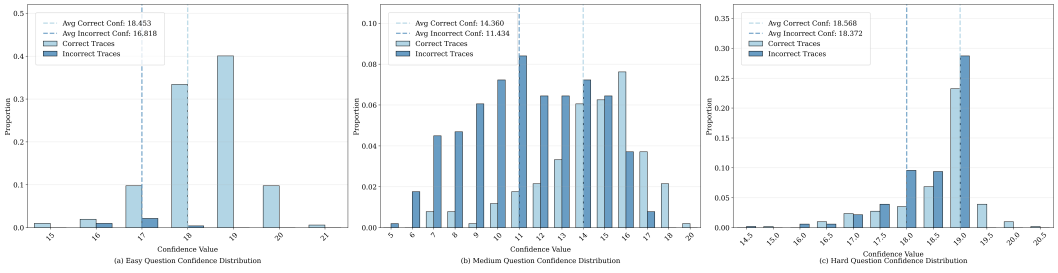


Figure 6: Analyzing the confidence distribution for problems of varying complexity.

Theorem B.1 (Sufficient Condition for Proxy Acceleration)**.** *Let N_{FrQ} and N_{CoW} be the continuous sample-size proxies at which the population signal-to-noise scores below reach $Z_\tau > 0$. Assume i.i.d. pairs (u_n, w_n), finite conditional second moments, and nonzero signed-evidence variances. Assume also:*

Assumption 1 (Solvability): $0.5 < p < 1$. ***Assumption 2 (Informative Weights):*** *Confidence weights are informative, meaning correct paths receive higher weights on average: $\mu_c > \mu_i$, where $\mu_c = \mathbb{E}[w \mid u_n = 1]$ and $\mu_i = \mathbb{E}[w \mid u_n = 0]$ as indicated in (Fu et al., 2025). Our experimental*

results of different complexity of MATH-500 (Hendrycks et al., 2021) in Figure 6 demonstrates this, showing a higher average confidence for correct paths.

Define the second moments $\mu_c^{(2)} = \mathbb{E}[w^2 \mid u_n = 1]$ *and* $\mu_i^{(2)} = \mathbb{E}[w^2 \mid u_n = 0]$. *Then the CoW proxy requires fewer samples,* $N_{CoW} < N_{FrQ}$, *if:*

$$\frac{p\mu_c - q\mu_i}{\sqrt{p\mu_c^{(2)} + q\mu_i^{(2)}}} > \frac{p-q}{\sqrt{4pq}}. \tag{14}$$

Proof. For FrQ, define the population signal-to-noise score from the Bernoulli mean and variance (Wickens, 1998):

$$Z_{FrQ}(N) = \sqrt{N}\frac{p-q}{\sqrt{4pq}}.$$

For CoW, we analyze the weighted process. Let $X_n = w_n(2u_n - 1)$ be the signed weight at step n. Then the expected weighted drift is $\mu_{CoW} = \mathbb{E}[X_n] = p\mu_c - q\mu_i$, and the variance is $\sigma_{CoW}^2 = \mathbb{E}[X_n^2] - \mu_{CoW}^2$. Using the inequality $\sigma_{CoW}^2 \le \mathbb{E}[X_n^2] = p\mu_c^{(2)} + q\mu_i^{(2)}$, we define $Z_{CoW}(N) = \sqrt{N}\mu_{CoW}/\sigma_{CoW}$, a population signal-to-noise proxy motivated by (Weiss, 2014), with lower bound:

$$\tilde{Z}_{CoW}(N) = \sqrt{N}\frac{p\mu_c - q\mu_i}{\sqrt{p\mu_c^{(2)} + q\mu_i^{(2)}}}.$$

Since $\mu_{CoW} > 0$ and $\sigma_{CoW}^2 \le \mathbb{E}[X_n^2]$, direct comparison of the denominators gives $Z_{CoW}(N) \ge \tilde{Z}_{CoW}(N)$. At the continuous crossing $Z_{FrQ}(N_{FrQ}) = Z_\tau$, condition equation 14 gives:

$$\tilde{Z}_{CoW}(N_{FrQ}) > Z_{FrQ}(N_{FrQ}) \ge Z_\tau.$$

Both scores increase as $\sqrt{N}$, so the continuous CoW crossing occurs strictly earlier: $N_{CoW} < N_{FrQ}$. This proves the proxy comparison, not a claim about random first-passage times or finite-budget correctness. □

Safety Brake Mechanism: Uniformly rescaling weights by $\epsilon > 0$ leaves the signal-to-noise ratio in equation 14 unchanged, but rescales the total evidence in the Beta posterior. In the special case $w_n = \epsilon \ll 1$ and a fixed limiting support fraction above 0.5, matching a given amount of unweighted evidence takes approximately $1/\epsilon$ times as many samples. This total evidence is not the importance-sampling effective sample size (Hammersley & Handscomb, 1964). The scale effect explains why proxy acceleration alone cannot guarantee earlier Beta stopping, and why raw weights above one do not provide an unconditional safety brake.

B.1.3 Empirical Verification

To examine the mechanism motivating Theorem B.1, we conducted Monte Carlo simulations across varying question difficulties: Easy Questions denotes level 1 of MATH500 (Hendrycks et al., 2021), Medium Questions denotes level 3, and Hard Questions denotes level 5.

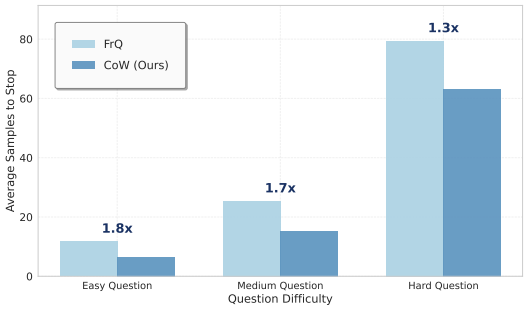


Figure 7: Empirical verification.

As shown in Figure 7, CoW demonstrates significant efficiency gains over FrQ when confidence information is reliable. These observations support the role of informative confidence in the tested settings, consistent with the signal-to-noise motivation of Theorem B.1. They do not establish uniform acceleration or correctness guarantees for adaptive sampling with potentially miscalibrated weights.

B.2 Intra-Path: Trend-Aware Pruning Analysis

B.2.1 Statistical Foundation of Pruning Threshold

We employ Tukey Fences (Tukey, 1977): $\tau_{\text{risk}} = Q_3 + 1.5 \cdot \text{IQR}$, applied to the instability score $\mathcal{R}$. This nonparametric threshold is robust to outliers and non-Gaussian distributions, addressing the skewed confidence distributions commonly observed in LLM reasoning. The interquartile range (IQR) provides a robust measure of dispersion, making the threshold adaptive to query difficulty without parametric assumptions.

B.2.2 Analysis of Failure Modes

Let $\mathbf{Z} \in \mathbb{R}^{L\times 2}$ be the phase-space trajectory with second-moment eigenvalues $\lambda_1 \geq \lambda_2$. Spectral analysis of confidence trajectories follows established methods for time series analysis.

Spectral structure alone cannot rule out a confidently wrong answer. To analyze the large-sample weighted vote, we introduce a conditional assumption on the retained-path distribution:

Assumption 3 (Inter-Path Diversity): For a fixed retained-path distribution with a finite answer set, i.i.d. samples, and finite expected weights, the correct answer u_{gt} has strictly greater expected weighted support than any single incorrect answer:

$$\mathbb{E}[w\,\mathbb{I}(\text{ans}(y) = u_{\text{gt}})] > \max_{u\neq u_{\text{gt}}} \mathbb{E}[w\,\mathbb{I}(\text{ans}(y) = u)]. \tag{15}$$

Diverse sampled reasoning paths motivate this assumption (Wang et al., 2023), but do not guarantee it. It is a condition for asymptotic weighted-plurality consistency, not for absolute-majority support, finite-budget correctness, or the adaptive early-stopping rule.

Proposition B.2 (Pruning Efficacy)**.** *The pruning rule and Assumption 3 imply the following conditional statements:*

1. ***Stochastic Hallucinations:*** *A path reaching the trend-analysis tier is pruned if $\mathcal{R} > \tau_{risk}$; stochasticity alone does not ensure this inequality.*
2. ***Decaying Hallucinations:*** *Negative alignment adds $\eta\, align^2$ to the risk, increasing it relative to the same spectral structure with nonnegative alignment, but not necessarily crossing the threshold.*
3. ***Self-Reinforcing Hallucinations:*** *Such paths may survive pruning. Under Assumption 3, the weighted argmax over an increasing number of retained samples converges almost surely to u_{gt}.*

Proof. The first two claims follow from the gating rule and the nonnegative velocity penalty; the Tukey fence (Tukey, 1977) supplies an outlier threshold, not a hallucination-detection probability. For the third claim, the strong law gives convergence of each normalized answer-weight sum to its expectation. The finite answer set and strict expected-support gap imply eventual selection of u_{gt}. This conditional aggregation result, related to consistency considerations in (Bernardo & Smith, 1994), does not transfer to a finite-budget adaptive stopping guarantee. □

B.3 Limitations and Broader Connections

Limitations and Future Work: DDC inherently relies on the strong correlation between confidence and correctness, offering no mechanism to circumvent systematic miscalibration (Guo et al., 2017), where in the model is confidently wrong across all reasoning paths.

The Bayesian termination in DDC is fundamentally driven by the contrast between the evidence strength of the dominant candidate,

$$\alpha_u = 1 + \sum w_i \cdot I\{\text{ans}_i = u\}, \tag{16}$$

and the aggregate counter-evidence,

$$\beta_u = 1 + \sum w_j \cdot I\{\text{ans}_j \neq u\}. \tag{17}$$

In scenarios where a model exhibits systematic miscalibration, it generates reasoning paths that are both identically incorrect and highly confident. Consequently, $\alpha_{u'}$ grows with each new sample, reflecting an accumulation of high-quality evidence in the eyes of the base model, while $\beta_{u'}$ remains near its initial Laplace prior value, as no dissenting paths are generated to challenge the false consensus. As supporting evidence accumulates, the posterior probability $P(p_{u'} > 0.5 \mid D)$ approaches one and can cross $\tau_{\text{stop}} = 0.95$, triggering an early exit as soon as initialization is complete.

Table 4: Computational cost comparison on MATH500 and AIME25. Width refers to the average number of generated paths per question; Token denotes the average token consumption for high-confidence but incorrect predictions.

Method	MATH500		AIME25	
	Width	Token	Width	Token
Qwen3-4B				
SC	512	0.81	512	0.96
DDC	**62**	**0.17**	**94**	**0.19**

The distributions in Figure 6 illustrate the association between confidence and correctness, while Table 4 examines confidently incorrect predictions. Systematically wrong consensus is an inherent boundary when internal confidence signals do not distinguish correctness. DDC can save further sampling on such a consensus, but it neither identifies the answer as wrong nor proves that the query is unresolvable. These are per-query savings, not an explicit mechanism for reallocating a shared budget between queries.

To further directly address this boundary, a promising avenue for future work is to integrate advanced confidence calibration techniques into the DDC framework. By incorporating post-hoc calibration or intrinsic epistemic uncertainty estimation to dynamically penalize or re-weight the evidence scores w_i, we anticipate that the framework can better decouple true consensus from systematic hallucinations, thereby further expanding the safe operating boundaries of inference-time scaling.

Method Complexity. DDC combines several interacting components, i.e., multi-granular confidence, confidence-weighted Bayesian termination, phase-space trend analysis, and Tukey gating, which raises a natural concern about complexity. Two factors keep this manageable. First, the auxiliary computation is negligible: a matrix-operation FLOP estimate is 1.25×10^{-7} of a 32B forward pass (Appendix C), while the measured driver-side analysis stack is 1.72% of wall-clock in the profiled configuration (Table 6). Second, the main comparison reuses one fixed hyperparameter set, including the design choice $\eta = 0.5$; the percentile and Tukey thresholds are computed from each query's initialization samples. The ablation *w/o Trend-Aware* isolates the cost of the most elaborate component, i.e., a 6.8-point drop on AIME25, indicating the trend layer earns its complexity; distilling the coupling into fewer components is a promising simplification for future work.

Generalization Beyond Short-Answer Math. DDC is domain-agnostic except for the answer-equivalence relation used during voting and consensus. To probe this, we replace numeric equality with an execution-based equivalence relation on HumanEval (Chen et al., 2021) code generation: the proof of concept uses $B_{\text{init}} = 8$ and $B = 16$, reaches a 91.5% pass rate over 164 problems, and terminates early on 48.2% of them. Equivalence uses execution signatures on visible examples; hidden tests are used only for evaluation. For open-ended generation, a semantic-entropy equivalence relation (Farquhar et al., 2024) can replace exact match, and agentic or multi-step tasks decompose into per-step reasoning decisions to which the same termination and pruning machinery applies. A full study across these regimes is left to future work.

Connections: DDC relates evidence reweighting to *Importance Sampling* (Hammersley & Handscomb, 1964), adaptive stopping to *Sequential Probability Ratio Testing* (Wald, 1945), and posterior construction to power-based Bayesian inference (Ibrahim et al., 2015). These are conceptual connections: DDC uses neither density-ratio weights nor a classical likelihood-ratio stopping test.

C More Experiment Results

C.1 Theoretical Complexity Analysis of DDC Operations

To address concerns regarding the computational overhead introduced by Trend-Aware Stratified Pruning and Bayesian Termination, we provide a formal complexity analysis comparing DDC's auxiliary operations with the standard LLM inference process.

Dimensionality of Auxiliary Operations: For each generation step t, DDC maintains $z_t \in \mathbb{R}^d$, where $d = 2$ represents normalized position and velocity. Over a window of L tokens, we construct $Z \in \mathbb{R}^{L\times 2}$.

- *Second-Moment Matrix Construction:* Computing $\Sigma = \frac{1}{L-1} Z^\top Z$ costs $O(Ld^2)$. Since $d = 2$, this is approximately $4L$ multiply–accumulate operations, excluding confidence extraction and data movement.
- *Eigendecomposition:* Performing eigendecomposition on a 2×2 matrix is an analytically solvable problem with a constant complexity of $O(d^3) = O(8)$.
- *Bayesian Update:* The calculation of the incomplete beta function $I_\gamma(\alpha, \beta)$ is a scalar operation performed once per path completion, with negligible cost.

Quantifying the Overhead Ratio: For a standard LLM with N parameters like 7B or 32B, the computation required for generating a single token is approximately $2N$ FLOPs.

For Qwen3-32B, one forward pass requires $\approx 64 \times 10^9$ FLOPs per token.

Counting each multiply–accumulate as one operation, the matrix construction at $L = 2048$ uses approximately 8×10^3 operations per checked token. This convention gives the illustrative ratio below; counting multiplication and addition separately doubles its numerator.

Overhead Ratio:

$$\frac{\text{matrix-operation estimate}}{\text{forward-pass FLOP estimate}} \approx 1.25 \times 10^{-7}.$$

This estimate concerns only the small matrix calculation, not the entire analysis or serving pipeline. Actual latency also depends on confidence extraction, scheduling, and batch utilization; token savings therefore do not translate one-to-one into wall-clock savings.

C.2 Analysis of Inference-Time Efficiency and Latency

Table 5 decomposes generation time into initialization, continuation, and their total. DDC is faster than DeepConf-Low in most reported settings, but not all: on Qwen3-4B/AMC23, the totals are 523.90s and 520.89s, respectively.

Superior Efficiency on Large-Scale Models: The advantage of DDC becomes more pronounced as the model size increases. For Qwen3-32B on the AIME25 benchmark: DeepConf-Low requires 24,912 seconds of total generation time, but DDC requires only 2,011 seconds, achieving a $12.4\times$ speedup. This dramatic reduction is primarily due to DDC's ability to trigger early termination much sooner than DeepConf-Low while maintaining comparable accuracy. For instance, for Qwen3-32B on the AIME25, DDC's FG time is 393s, whereas DeepConf-Low's is 22850s, a $58\times$ faster finalization.

The Warmup-to-Final Trade-off: In some settings, such as Qwen3-1.7B, DDC has higher warmup generation time than DeepConf-Low. This cost is often offset by shorter continuation, although warmup differences alone do not identify the cause and the gain is not universal.

Robustness Across Task Difficulty: On the most challenging AIME25 benchmark, DDC not only saves time but also significantly improves accuracy from 73.3% to 82.1% on Qwen3-4B. This is consistent with a useful joint accuracy–cost trade-off, but aggregate accuracy alone does not identify which incorrect paths were filtered.

The measured generation-time reductions support DDC's efficiency in these settings. The nonzero analysis and batching costs are examined separately below.

Table 5: Latency comparison between DeepConf-Low and DDC. Acc. is reported accuracy; WG, FG, and TG are mean per-question warmup, continuation, and total generation time in seconds, with TG=WG+FG up to rounding. These generation timings are distinct from the serving-profile wall-clock breakdown. Best results are in bold.

Method	**AMC23**				**AIME24**				**AIME25**			
	Acc. ↑	WG ↓	FG ↓	TG ↓	Acc. ↑	WG ↓	FG ↓	TG ↓	Acc. ↑	WG ↓	FG ↓	TG ↓
Qwen3-0.6B												
D-L	65.0	**90.92**	1098.35	1189.26	33.0	**116.20**	1571.14	1687.34	30.0	**116.20**	1571.14	1687.34
DDC	**70.0**	212.03	**328.43**	**540.46**	**36.7**	386.31	**618.89**	**1005.20**	**36.7**	305.59	**538.11**	**843.71**
Qwen3-1.7B												
D-L	92.5	**97.05**	1470.12	1567.17	**76.7**	**131.13**	2141.64	2272.76	**53.3**	**127.83**	2042.16	2169.99
DDC	**95.0**	447.05	**108.76**	**555.81**	73.3	748.99	**531.82**	**1280.81**	50.0	808.66	**487.71**	**1296.37**
Qwen3-4B												
D-L	100.0	**494.25**	26.65	**520.89**	83.1	880.10	7721.52	8601.62	73.3	1093.82	13483.31	14577.12
DDC	**100.0**	506.28	**17.62**	523.90	**83.3**	**811.43**	**71.00**	**882.43**	**82.1**	**857.79**	**149.24**	**1007.03**
Qwen3-8B												
D-L	96.7	**469.15**	5058.45	5527.59	**80.4**	1056.11	10610.01	11666.12	**76.7**	1374.06	15233.55	16607.62
DDC	**97.5**	492.25	**56.19**	**548.45**	80.0	**833.45**	**130.89**	**964.35**	73.3	**940.04**	**226.82**	**1166.86**
Qwen3-32B												
D-L	97.3	708.83	7610.27	8319.10	89.4	1584.69	15915.42	17500.11	**80.2**	2061.38	22850.73	24912.11
DDC	**97.5**	**706.61**	**36.89**	**743.51**	**93.3**	**1296.13**	**197.42**	**1493.55**	79.7	**1617.74**	**393.36**	**2011.10**

End-to-End Wall-Clock Decomposition. We profile Qwen3-4B on 30 AIME25 questions using one NVIDIA L20Z, $B_{\text{init}} = 16$, and DDC budget $B = 64$. Table 6 separates warmup, continuation, CPU analysis, and scheduling in this configuration; its percentages are not assumed invariant to budget.

Table 6: End-to-end wall-clock decomposition of DDC on a single GPU (Qwen3-4B, AIME25). The driver-side analysis stack, bundling multi-granular confidence, the Beta-CDF stopping check, phase-space trend analysis, and Tukey thresholding, accounts for only 1.72% of total wall-clock.

Wall-clock component	Share
Warmup generation	94.6%
Continued generation	3.3%
CPU analysis stack	1.72%
Scheduling residual	0.34%

The 1.72% is the cumulative CPU analysis time: approximately 174.2s over the run, dominated by confidence and trend computation. The cumulative Beta-CDF, Tukey-fence, and weighted-vote times are 0.002s, 0.0015s, and 0.007s, respectively, not per-call costs. Table 7 reports the associated serving behavior.

Table 7: Serving profile on one L20Z (Qwen3-4B/AIME25): full-batch SC@16 versus DDC@64. Budgets differ, so this is not a matched-budget speedup comparison. The measured wall-clock excess over the ideal time at SC throughput is 25.4% of DDC wall-clock.

Serving metric	SC (full batch)	DDC
Throughput (tok/s)	1,270	947
Peak GPU memory (GB)	80.2	80.7
GPU utilization	92.7%	88.8%

Table 8: Performance across initialization sizes. Acc. is accuracy (%); WT, FT, and TT denote initialization, continuation, and total tokens ($\times 10^7$). Avg. averages only AMC23, AIME24, and AIME25, unlike the five-task average in the main table.

B_{init}	AMC23				AIME24				AIME25				Avg	
	Acc. ↑	WT ↓	FT ↓	TT ↓	Acc. ↑	WT ↓	FT ↓	TT ↓	Acc. ↑	WT ↓	FT ↓	TT ↓	Acc. ↑	TT ↓
Qwen3-4B														
8	97.5	**0.26**	0.54	0.80	80.0	**0.34**	1.12	1.46	73.3	**0.42**	2.34	2.77	83.6	1.68
16	**100.0**	0.39	**0.12**	**0.51**	**83.3**	0.69	0.47	**1.16**	**82.1**	0.75	**0.71**	**1.46**	**88.5**	**1.04**
32	100.0	1.04	0.00	1.04	80.0	1.40	**0.35**	1.75	80.0	1.77	1.37	3.14	86.7	1.98
Qwen3-8B														
8	**100.0**	**0.29**	0.38	**0.67**	**83.3**	**0.36**	1.15	1.51	**76.7**	**0.43**	2.27	2.69	**86.7**	1.62
16	97.5	0.57	0.44	1.01	80.0	0.71	**0.57**	**1.29**	73.3	0.86	1.31	**2.17**	83.6	**1.49**
32	100.0	1.14	**0.18**	1.32	80.0	1.44	0.71	2.15	76.7	1.79	**0.89**	2.68	85.6	2.05
DeepSeek-R1-0528-Qwen3-8B														
8	100.0	0.38	**0.38**	**0.76**	83.3	**0.36**	1.15	**1.51**	83.3	**0.52**	1.70	**2.22**	88.9	**1.50**
16	**100.0**	0.80	**0.00**	0.80	**86.7**	1.01	0.92	1.93	**83.3**	1.07	1.30	2.37	**90.0**	1.70
32	100.0	0.92	0.00	0.92	76.7	1.48	**0.71**	2.19	80.0	2.12	**1.25**	3.37	85.6	2.16

Table 9: L(x) denotes a window of x tokens. FT and TT are continuation and total tokens ($\times 10^7$); Avg. averages the three competition benchmarks.

Window size	AMC23			AIME24			AIME25			Avg		
	Acc. ↑	FT ↓	TT ↓	Acc. ↑	FT ↓	TT ↓	Acc. ↑	FT ↓	TT ↓	Acc. ↑	FT ↓	TT ↓
Qwen3-4B												
L(512)	100.0	**0.03**	0.54	80.0	**0.08**	**0.78**	76.7	**0.15**	**0.98**	85.6	0.09	**0.77**
L(1024)	100.0	0.12	0.63	80.0	0.30	1.00	73.3	0.48	1.32	84.4	**0.30**	0.98
L(2048)	**100.0**	0.12	**0.51**	**83.3**	0.47	1.16	**82.1**	0.71	1.46	**88.5**	0.43	1.04
Qwen3-8B												
L(512)	97.5	**0.00**	**0.55**	80.0	**0.10**	**0.83**	**80.0**	**0.21**	**1.06**	**85.8**	**0.10**	**0.81**
L(1024)	97.5	0.15	0.71	**83.3**	0.23	0.94	76.7	0.54	1.40	85.8	0.31	1.02
L(2048)	**97.5**	0.44	1.01	80.0	0.57	1.29	73.3	1.31	2.17	83.6	0.77	1.49
DeepSeek-R1-0528-Qwen3-8B												
L(512)	100.0	0.00	0.76	86.7	0.10	1.09	83.3	**0.18**	**1.23**	90.0	**0.09**	**1.03**
L(1024)	100.0	0.00	0.77	90.0	0.46	0.98	83.3	0.46	1.52	**91.1**	0.31	1.09
L(2048)	**100.0**	**0.00**	0.80	86.7	0.92	1.93	**83.3**	1.30	2.37	90.0	0.74	1.70

The throughput-based 25.4% wall-clock gap already includes effects such as analysis and scheduling; it is not additive with the 1.72% CPU share or an isolated causal estimate of pruning overhead. This profile illustrates why token reduction alone does not determine wall-clock speedup.

C.3 Sensitivity to Init Budget

We investigate the sensitivity of performance to the initialization budget B_{init}, which controls the number of reasoning paths generated during the initialization phase. As summarized in Table 8, the choice of B_{init} exhibits a clear trade-off between statistical reliability and computational cost across different models. For Qwen3-4B, this setting yields 88.5% average accuracy while consuming only 1.04×10^7 tokens. For Qwen3-4B, $B_{\text{init}} = 8$ has lower reported accuracy, while $B_{\text{init}} = 32$ costs more without an accuracy gain over 16. These trends are not universal: Qwen3-8B has higher accuracy with 8 or 32 than with 16. This demonstrates that a calibrated initial exploration phase is helpful for establishing reliable baselines.

C.4 Effect of Window Granularity

We analyze the window size L used to compute group confidence. Results in Table 9 reveal that larger windows $L = 2048$ capture longer reasoning paths, leading to better performance on complex tasks like AIME25 82.1% vs. 76.7% with $L = 512$ for Qwen3-4B. However, smaller windows $L = 512$ can be more efficient for simpler problems. Our analysis confirms that longer context windows help distinguish genuine reasoning trends from fluctuations, improving the reliability of quality assessments.

Together, these analyses validate our parameter choices $B_{\text{init}} = 16$, $L = 2048$ as providing a reliable default configuration that balances accuracy, efficiency, and robustness across diverse reasoning tasks and models.

C.5 Scaling behavior

Figure 8 reports accuracy and token consumption versus voting size across models and datasets. Voting-size analysis of saved trajectories is distinct from an online attempt-budget sweep: a voting size below B_{init} does not represent online DDC with a smaller initialization.

C.6 Calibration of the Confidence Signal

DDC uses confidence as a quality proxy, not a calibrated correctness probability. On an 11-question AIME25 subset with 16 saved Qwen3-4B paths per question (176 paths), the reported AUROC is 0.74. The reported 5-fold cross-validated Platt-scaled ECE is 0.078; raw unnormalized scores are not probabilities. Correct paths have higher mean scores on this subset, providing local evidence for Assumption 2, not a universal guarantee. Separately, the tested power transformations ($w \to w^{0.5}$ and $w \to w^2$) caused no vote changes in the saved-pool tests. This observation does not imply invariance: monotone transformations can change weighted sums, phase-space geometry, Tukey thresholds, and Beta concentration. The systematic-miscalibration boundary in Appendix B.3 therefore remains.

C.7 Additional Adaptive, Stacked, and Verifier-Guided Baselines

Table 10 compares reported Qwen3-4B results for MUR (Yan et al., 2025) (depth only), RASC (Kim et al., 2026) and ESC (Li et al., 2024) (width only), and their post-hoc stackings with DeepConf (Fu et al., 2025) and MUR. A verifier-guided ReasonFlux-PRM-7B (Zou et al., 2025) baseline, which additionally uses an external 7.62B verifier, is reported separately in the text below.

Table 10: Adaptive and stacked baselines on Qwen3-4B, with maximum path budget $N = 512$. Accuracy (%) and tokens ($\times 10^7$). Main-table SC and DDC reference values are repeated unchanged.

Method	Adaptivity	AIME25		GPQA-d	
		Acc.	Tokens	Acc.	Tokens
MUR	depth only	16.7	0.015	40.9	0.020
RASC	width only	76.7	0.45	59.1	1.89
ESC	width only	76.7	8.12	58.6	12.0
ESC + DeepConf	stacking, post hoc	76.7	8.12	58.6	12.0
ESC + MUR	stacking, post hoc	76.7	5.1	59.1	11.4
SC	none	66.5	41.0	48.5	80.4
DDC	coupled	**82.1**	1.5	**59.1**	13.7

Online coupling and post-hoc stacking. ESC-style stopping plateaus at 76.7% on AIME25 even as its token use grows approximately $13.5\times$ (from 0.6 to 8.12×10^7), and post-hoc DeepConf weighting flips no decisions relative to ESC. This indicates that post-hoc weighting alone does not improve ESC; DDC's advantage instead comes from coupling confidence-based termination with per-path quality control during generation. The reported DDC reference is 82.1% at roughly $1/5$ of ESC's token count.

Verifier-guided selection. On AIME25, Best-of-N and PRM-weighted voting with ReasonFlux-PRM-7B improve over SC ($66.5 \to 73.3\%$ at 47.6×10^7 tokens) but, by construction, cost more than

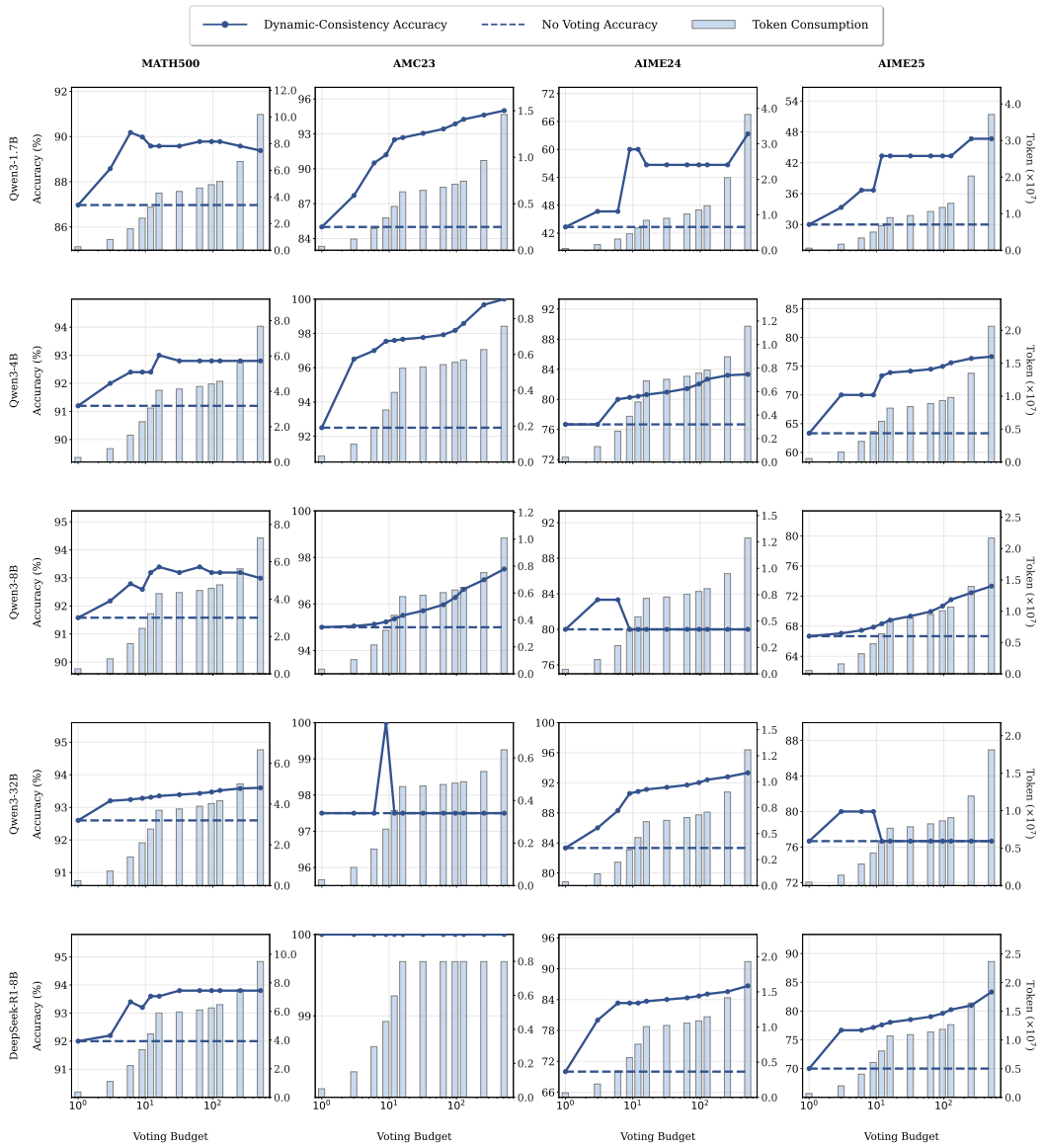


Figure 8: Scaling behavior of different models.

SC: they generate all 512 paths and then pay the verifier forward passes. DDC exceeds both while consuming about $1/30$ of their tokens, and a PRM remains composable with DDC since its step score can serve as our intra-path quality channel.

An honest boundary on GPQA. RASC matches the reported DDC accuracy on GPQA-diamond (59.1%) at fewer tokens. On AIME25, its reported accuracy is 76.7%, 5.4 points below DDC. Thus DDC does not dominate every baseline in every cell. Neither similar accuracies nor the absence of a correct answer in a saved pool establishes a model-capability ceiling.

C.8 Failure-Case Taxonomy

We distinguish three observational failure classes on Qwen3-4B (Table 11). We inspect the reasoning paths actually saved for each query and report each class as a share of the analyzed errors, without assuming a fixed per-query path count.

Table 11: Failure taxonomy on Qwen3-4B, computed over the reasoning paths actually saved per query and reported as a share of the analyzed errors.

Failure class	AIME25	GPQA-d
No correct answer among the saved paths	80.0%	56.1%
Correct answer saved but outvoted	20.0%	31.7%
Wrong early stop while a correct answer was available	0.0%	12.2%

The dominant class also affects Self-Consistency: when no correct path exists among the samples analyzed, no voting or allocation policy can recover it; we do not claim this proves a model-capability ceiling without further exploration. Premature lock-in, where a correct answer was available but DDC committed early to a wrong one, is 0% of analyzed AIME25 errors and 12.2% on GPQA. As a complementary diagnostic, wrong early-stops terminate within a few paths and are therefore cheap, consistent with the low token cost of high-confidence incorrect predictions in Table 4. The exact denominators (evaluated queries and saved paths) vary by dataset and are omitted here to avoid conflating offline path-pool statistics with the online token budget.

Where DDC trails the strongest baseline. At full budget, DDC is slightly below the best baseline on a few cells, i.e., Qwen3-8B AIME25 (73.3% vs. 78.5%), Qwen3-32B AIME25 (79.7% vs. 80.2%), and DeepSeek-R1 on average (86.7% vs. 87.1%). The comparison methods in these examples are D-H or D-L, not uniformly SC. Late recovery is one possible explanation, but the aggregate results do not establish that cause. Early stopping and pruning can both trade accuracy for compute, so the claim is a favorable overall trade-off rather than per-cell optimality.

C.9 Robustness to Temperature, Threshold, Confidence Metric, and Model Family

Beyond initialization and window size, we fix the decoding temperature and stopping threshold and examine the confidence-metric choice and model family. All main results use $T = 0.6$ and $\tau_{\text{stop}} = 0.95$; under this configuration DDC reaches 82.1% on Qwen3-4B/AIME25 at 1.5×10^7 tokens.

Temperature. A single decoding temperature $T = 0.6$ is used for every method, so observed differences isolate the scaling strategy rather than the sampler.

Stopping threshold. τ_{stop} is a compute knob: a higher value demands stronger weighted consensus before halting and thus tends to sample more paths, trading accuracy for compute. We use $\tau_{\text{stop}} = 0.95$ throughout.

Confidence metric and channel assignment. In saved-pool voting tests, top-1 confidence, negative entropy, and self-certainty agree on $96.7-100\%$ of decisions, so the consensus outcome is insensitive to this particular proxy. Negative entropy is an alternative metric, distinct from the top-k negative mean log-probability defined in Section 4, and this agreement is measured over saved pools rather than online generation.

Model family. Beyond the four Qwen3 backbones and the DeepSeek distillation in Table 1, we add GPT-OSS-20B (OpenAI, 2025), an additional model family (Table 12). DDC matches or exceeds both DeepConf variants at $9-35\times$ fewer tokens, confirming the mechanism is not tied to the Qwen family.

Table 12: GPT-OSS-20B: accuracy (%) and tokens ($\times 10^7$). DDC matches or exceeds both DeepConf variants at far fewer tokens.

Dataset	D-L		D-H		**DDC**	
	Acc.	Tok.	Acc.	Tok.	Acc.	Tok.
AIME24	96.7	11.1	96.7	30.7	**96.7**	**1.2**
AIME25	94.0	12.1	94.8	31.8	**96.7**	**0.9**

Algorithm 1 Initialization Phase

Require: Query x, Init Budget B_{init}, Window Size L
Ensure: $\mathcal{S}$, $\mathcal{C}_{\text{local}}$, $\mathcal{C}_{\text{global}}$, $\mathcal{R}_{\text{vals}}$, τ_{pass}, τ_{drop}, τ_{risk}, α, β
Initialize empty multisets $\mathcal{S}, \mathcal{C}_{\text{local}}, \mathcal{C}_{\text{global}}, \mathcal{R}_{\text{vals}}$
for $i = 1$ to B_{init} **do**
  Generate $y_i = (y_1, \ldots, y_T) \sim P_\theta(\cdot \mid x)$ until EOS or $T_{\max}$
  Compute $C_t^l = P_t(1)$ and $C_t^g = -\frac{1}{k}\sum_{j=1}^{k} \log P_t(j)$ for $t \in [1, T]$
  $\forall G_t \in \mathcal{G}(y_i) : C_{G_t}^l \leftarrow \frac{1}{|G_t|}\sum_{j \in G_t} C_j^l, \quad C_{G_t}^g \leftarrow \frac{1}{|G_t|}\sum_{j \in G_t} C_j^g$
  $C_{\text{path}}(y_i) \leftarrow \min_{G_t} C_{G_t}^g$
  Construct matrices $\mathbf{Z}_t$ from window $\{C_j^g\}_{j \in G_t}$
  Compute $\mathcal{R}_t$ using Eq. equation 9 and its degenerate-window convention
  Append all $C_{G_t}^l$, $C_{G_t}^g$, and $\mathcal{R}_t$ to their respective multisets, preserving repetitions
  if $\text{ans}(y_i)$ is extractable **then**
    Append y_i to $\mathcal{S}$
  end if
end for
// Calibrate Query-Specific Thresholds
$\tau_{\text{pass}} \leftarrow \text{Percentile}(\mathcal{C}_{\text{local}}, 90)$
$\tau_{\text{drop}} \leftarrow \text{Percentile}(\mathcal{C}_{\text{global}}, 20)$
$\tau_{\text{risk}} \leftarrow Q_3(\mathcal{R}_{\text{vals}}) + 1.5 \cdot \text{IQR}(\mathcal{R}_{\text{vals}})$
Initialize Bayesian Termination
$\alpha_u \leftarrow 1, \beta_u \leftarrow 1$ for all $u \in \mathcal{U}$
for $y \in \mathcal{S}$ **do**
  $u' \leftarrow \text{ans}(y)$; $w \leftarrow C_{\text{path}}(y)$
  $\alpha_{u'} \leftarrow \alpha_{u'} + w; \quad \forall u \neq u', \beta_u \leftarrow \beta_u + w$
end for
return $\mathcal{S}$, $\mathcal{C}_{\text{local}}$, $\mathcal{C}_{\text{global}}$, $\mathcal{R}_{\text{vals}}$, τ_{pass}, τ_{drop}, τ_{risk}, α, β

Algorithm 2 Confidence-Weighted Bayesian Termination

Require: Advantage counts $\alpha = \{\alpha_u\}$, disadvantage counts $\beta = \{\beta_u\}$
Ensure: Boolean indicating whether to terminate
$u^* \leftarrow \arg\max_u \alpha_u$
$p_{\text{majority}} \leftarrow 1 - I_{0.5}(\alpha_{u^*}, \beta_{u^*})$
if $p_{\text{majority}} > 0.95$ **then**
  return true
else
  return false
end if

D ALGORITHM

The complete workflow of Dual-Dimensional Consistency is presented in Algorithm 1 (Initialization Phase), Algorithm 2 (Bayesian Termination), Algorithm 3 (Stratified Generation), and Algorithm 4 (Main Framework). The main framework first performs initialization, then generates further paths while checking termination, and finally aggregates results. B counts all generation attempts, including pruned ones; $\mathcal{S}$ is a multiset of retained paths with extractable answers. Generation ends at EOS or the length limit; length-capped outputs are not assumed to be complete. If no answer-bearing path is retained, the method abstains.

Algorithm 3 Trend-Aware Stratified Generation

Require: Query x, Thresholds $\tau_{\text{pass}}, \tau_{\text{drop}}, \tau_{\text{risk}}$, Window L
Ensure: Unpruned path y (possibly length-capped) or NULL

$y \leftarrow \emptyset$
for $t = 1$ to $T_{\max}$ **do**
  $y_t \sim P_\theta(\cdot \mid x, y_{<t})$
  Append y_t to y and record $C_t^l = P_t(1)$ and C_t^g
  if y_t is EOS **then**
    return y
  end if
  if $t < L$ **then**
    continue
  end if
  Define window indices $G_t = \{t - L + 1, \ldots, t\}$
  // Tier 1: Local Confidence Check
  $C_{G_t}^l \leftarrow \frac{1}{L} \sum_{j \in G_t} C_j^l$
  if $C_{G_t}^l > \tau_{\text{pass}}$ **then**
    continue
  end if
  // Tier 2: Global Confidence Check
  $C_{G_t}^g \leftarrow \frac{1}{L} \sum_{j \in G_t} C_j^g$
  if $C_{G_t}^g < \tau_{\text{drop}}$ **then**
    return NULL
  end if
  // Tier 3: Trend-Aware Analysis
  Construct $\mathbf{z}_j = [\hat{C}_j^g, \Delta\hat{C}_j^g]^\top$ for $j \in G_t$
  $\mathbf{Z} \leftarrow [\mathbf{z}_{t-L+1}, \ldots, \mathbf{z}_t]^\top$
  Compute $\lambda_1, \lambda_2, \vec{v_1}$ from $\frac{1}{L-1}\mathbf{Z}^\top\mathbf{Z}$
  Compute align and $\mathcal{R}$ using Eq. equation 9 and its degenerate-window convention
  if $\mathcal{R} > \tau_{\text{risk}}$ **then**
    return NULL
  end if
end for
return y

Algorithm 4 Dual-Dimensional Consistency Framework

Require: Query x, budgets $B \geq B_{\text{init}} \geq 1$, Window Size $L \geq 2$
Ensure: Final Answer $\hat{y}$

// Initialization Phase
Execute Algorithm 1 with inputs x, B_{init}, L to obtain
    $\mathcal{S}$, $\mathcal{C}_{\text{local}}$, $\mathcal{C}_{\text{global}}$, $\mathcal{R}_{\text{vals}}$, τ_{pass}, τ_{drop}, τ_{risk}, α, β
$n_{\text{attempt}} \leftarrow B_{\text{init}}$
while $n_{\text{attempt}} < B$ **do**
    terminate $\leftarrow$ ConfidenceWeightedBayesianTermination(α, β)
    if terminate **then**
        break
    end if
    $y_{\text{new}} \leftarrow$ TrendAwareStratifiedGeneration$(x, \tau_{\text{pass}}, \tau_{\text{drop}}, \tau_{\text{risk}}, L)$
    $n_{\text{attempt}} \leftarrow n_{\text{attempt}} + 1$
    if $y_{\text{new}} \neq$ NULL and $\text{ans}(y_{\text{new}})$ is extractable **then**
        Append y_{new} to $\mathcal{S}$
        $u' \leftarrow \text{ans}(y_{\text{new}})$; $w \leftarrow C_{\text{path}}(y_{\text{new}})$
        $\alpha_{u'} \leftarrow \alpha_{u'} + w$
        $\forall u \neq u', \beta_u \leftarrow \beta_u + w$
    end if
end while
if $\mathcal{S} = \emptyset$ **then**
    return abstention
end if
return $\arg\max_u \sum_{y \in \mathcal{S}} \mathbb{I}\{\text{ans}(y) = u\} \cdot C_{\text{path}}(y)$