PROBABILITY IS NOT ENOUGH: EXPLORING AND COUNTING DIVERGENT TOKENS FOR REASONING UNCERTAINTY QUANTIFICATION IN LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

As the chain-of-thought reasoning capabilities of large language models improve, evaluating and calibrating their reasoning confidence is becoming increasingly important for quantifying the uncertainty of their answers. Current methods for estimating the confidence of large language models are generally based on probabilities of selected key tokens, but the underlying mechanism remains unclear. Our pilot study finds that replacing selected token probabilities with coarse substitutes can also improve calibration, motivating us to further explore effective signals of model confidence. We introduce Divergent Token Confidence (DTC), a framework that estimates confidence by counting tokens at which two models strongly disagree during decoding. DTC identifies these divergent tokens using the Jensen–Shannon divergence between next-token distributions evaluated along the same reasoning trajectory. We find that their count is almost negatively associated with answer accuracy, thereby can serve as a simple yet effective signal for uncertainty quantification. DTC supports both white-box and black-box evaluation using auxiliary models, without explicit training and affecting the generation process. Experiments across multiple model families and six mathematical benchmarks demonstrate improved calibration over the considered probability-based and verbalized-based baselines. Under white-box evaluation, the count-only estimator achieves an average expected calibration error of 13.0%, compared with 32.7%–42.4% for standard full-sequence confidence methods. In black-box settings, it also improves calibration over the original verbalized scores. For example, mean expected calibration error falls from 32.1%–40.2% to 13.7%–16.3% on DeepSeek-V3.2. These findings provide new insights for improving reasoning uncertainty quantification in large language models. We release the code at anonymous repository

1 INTRODUCTION

Large language models have shown strong capability to solve complex tasks through Chain-of-Thought (CoT) reasoning (Wei et al., 2022; OpenAI, 2024; Guo et al., 2025). However, a reasoning trajectory that yields a correct answer may still contain unreliable intermediate decisions, while one that yields an incorrect answer may still be logically reliable (Wei et al., 2022; Bao et al., 2025; Zhou et al., 2026). Therefore, estimating the confidence of a model's reasoning trajectory is becoming increasingly important. First, a mismatch between model confidence and its answers can amplify the model's unreliability, thereby limiting its deployment in safety-critical scenarios (Clusmann et al., 2023). Second, understanding reasoning confidence can help us determine whether we need to rely on the outputs of models. When a reasoning path is identified as unreliable, the model can abstain from answering (Madhusudhan et al., 2025), or it can be deferred to human review (Devic et al., 2025). Finally, it has been shown that confidence itself can also be leveraged to improve the model's own reasoning performance (Fu et al., 2026).

We refer to this confidence estimation task as *Reasoning Uncertainty Quantification*: assigning a confidence score to each query and its reasoning path to reflect the reliability of the path and its final answer (Liu et al., 2025; Zhang & Zhang, 2025). Existing methods typically aggregate token probabilities over the entire trajectory, but the resulting scores can be overconfident (Orgad et al.,

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.

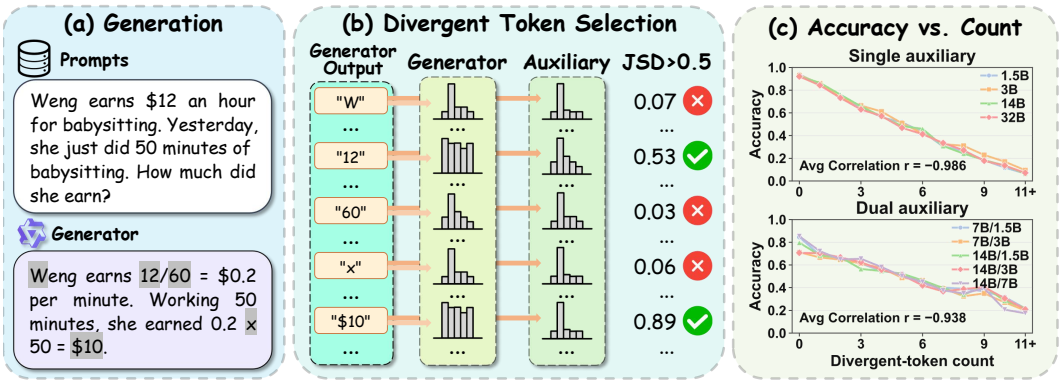

Figure 1: Divergent-token selection and its relationship to path accuracy. (a) A generator produces a reasoning trajectory. (b) A token is selected when the JSD between two models' next-token distributions exceeds θ. (c) Path accuracy decreases as the number of divergent tokens increases under single- and dual-auxiliary probing.

2025; Li et al., 2025; Zhang & Zhang, 2025). To mitigate this overconfidence, Li et al. (2025) propose Uncertainty Quantification with Attention Chain (UQAC), which selects answer-related CoT tokens and multiplies their probabilities to obtain a path-level confidence score. This design suggests that the probabilities of answer-related tokens provide useful information for estimating confidence in answer correctness. However, its product score depends jointly on these probabilities and the selected-token count, leaving it unclear whether the calibration gains come from the selected probabilities or the count.

To test whether the selected probabilities are necessary for these calibration gains, we keep UQAC's selected token set fixed and replace the selected probabilities with either the trajectory's mean token probability or a constant shared across trajectories from the same model on a given dataset. Both replacements generally improve calibration in our case study, suggesting that the selected tokens' specific probabilities may not be necessary for these gains. In particular, replacing the probabilities with a shared constant makes the score depend only on the number of selected tokens, motivating us to examine token count as a calibration signal (Section 3.3).

To obtain the token count signal that reflects reasoning path reliability, we draw on inter-model disagreement as an uncertainty signal (Kruse et al., 2025; Sun et al., 2024). We hypothesize that *unreliable reasoning paths contain more tokens at which models strongly disagree*. We measure token-level disagreement using the Jensen–Shannon divergence (JSD) between two models' next-token distributions and define tokens whose divergence exceeds a threshold θ as divergent tokens. Their count summarizes the frequency of strong disagreement along the reasoning path and is negatively associated with path accuracy in our experiments (Figure 1).

We therefore introduce **Divergent Token Confidence (DTC)**, which converts divergent-token count into path-level confidence. DTC provides two estimators: $\mathrm{DTC}_{\mathrm{lin}}$ maps the count directly to confidence, while $\mathrm{DTC}_{\mathrm{prod}}$ combines the count with a full-sequence probability score. The same procedure can be applied in white-box settings using the generator's information and in black-box settings using auxiliary models on the generated trajectory, without requiring access to the generator's logits. Across multiple model families and mathematical benchmarks, DTC improves calibration over full-sequence likelihood-based, verbalized-confidence, and UQAC baselines (Sections 5.1 and 5.2). DTC also reduces overconfidence in verbalized scores on existing trajectories (Section 6).

2 Related Work

Probability-based confidence and verbalized confidence. In white-box uncertainty quantification, token probabilities or predictive uncertainty from the generating model are commonly used to construct response-level confidence scores. Sequence-level uncertainty estimators typically aggregate token-level signals over a generated response (Malinin & Gales, 2021). Relevance-aware methods account for differences in how tokens contribute to meaning or the final answer, weighting

these signals by token relevance, sentence relevance, or contextual information (Duan et al., 2024; Bakman et al., 2024; Lin et al., 2024). For responses that include CoT reasoning, CoT-UQ and UQAC further use the relevance of tokens in the reasoning chain to the final answer to estimate response-level uncertainty (Zhang & Zhang, 2025; Li et al., 2025). In black-box settings, verbalized confidence is a common approach to confidence estimation (Wang & Zhang, 2026). Such methods either elicit confidence through additional prompts after a response has been generated or ask the model to report confidence or a distribution over candidate answers alongside its CoT and final answer (Tian et al., 2023; Wang et al., 2025). The latter couples reasoning with confidence elicitation: the choice of prompt may alter the CoT trajectory and lead to lower accuracy or overconfidence in mathematical reasoning (Wang et al., 2025). Our work reveals their limitations and investigates the novel divergent-token count as a confidence signal in both white-box and black-box settings.

Cross-model signals for confidence estimation. Another line of work uses multiple models, base models, or model perturbations to estimate the reliability of generated responses. MUSE and CrossCheckGPT use multi-model consensus or cross-system consistency to improve confidence estimation (Kruse et al., 2025; Sun et al., 2024). BaseCal uses signals from base models to calibrate the confidence of post-trained models (Tan et al., 2026). More closely related to our work are methods that estimate reasoning uncertainty from token entropy. Some aggregate the generating model's token entropy along a CoT trajectory; others use randomly perturbed models or additional models and aggregate their token entropy along the same trajectory to estimate uncertainty over the reasoning trajectory (Zhang et al., 2026; Gorbett & Jana, 2026). These studies primarily focus on response-level uncertainty or its ranking performance, whereas our focus is on a bounded score that can be directly interpreted as confidence and used for calibration.

3 Preliminaries and Pilot Study

3.1 Reasoning Uncertainty and Calibration

The *prediction uncertainty* (or query uncertainty) $U(x)$ characterizes uncertainty over answers to a query x before conditioning on a particular realized reasoning path, and is often estimated by sampling multiple answers to the same query (Kuhn et al., 2023). In this work, we study *reasoning uncertainty* $U(x, r)$, the uncertainty in final-answer reliability conditioned on both the query x and the realized reasoning process r (Liu et al., 2025; Zhang & Zhang, 2025).

Given the query x, a generating model $\mathcal{G}$ produces a trajectory $\tau = (\tau_1, \ldots, \tau_T)$ of T tokens, consisting of a CoT path r and a final answer y; a reference answer or evaluator provides the binary correctness label $z \in \{0, 1\}$. We operationalize $U(x, r)$ through a calibrated confidence $C(y, x, r) \in [0, 1]$ that estimates $P(z{=}1 \mid x, r, y)$, with higher confidence indicating lower reasoning uncertainty. Uncertainty is perfectly calibrated if

$$P\big(z{=}1 \mid y, x, r\big) = q, \text{ s.t. } C(y, x, r) = q. \tag{1}$$

Expected calibration error (ECE; Guo et al., 2017) is used to quantify the degree of miscalibration: we partitions N scored trajectories into M equal-width confidence bins $\{B_1, \ldots, B_M\}$ and then calculates the absolute gap between average accuracy $\mathrm{acc}(B_m)$ and confidence $\mathrm{conf}(B_m)$:

$$\mathrm{ECE} \;=\; \sum_{m=1}^{M} \frac{|B_m|}{N} \big|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\big|. \tag{2}$$

Lower ECE indicates better calibration. Unless noted, we report ECE with $M{=}20$.

3.2 Sequence Confidence from Token Probabilities

Standard full-sequence confidence scores aggregate token probabilities over the entire trajectory. One such score is the length-normalized sequence likelihood (NSL; Malinin & Gales, 2021). Given the generator's token probabilities $p_t = P_{\mathcal{G}}(\tau_t \mid x, \tau_{<t})$, we have

$$C_{\mathrm{NSL}}(\tau, x) \;=\; \Big(\prod_{t=1}^{T} p_t\Big)^{1/T} \;=\; \exp\Big(\frac{1}{T}\sum_{t=1}^{T} \log p_t\Big). \tag{3}$$

Another similar full-sequence confidence score is the mean token probability (Orgad et al., 2025), denoted as $C_{\text{mean}} = \frac{1}{T}\sum_{t=1}^{T} p_t$, which could be overconfident (Zhang & Zhang, 2025).

Recent work argue that not every token is equally diagnostic of answer correctness and accordingly assigns each token a relevance weight w_t to form a relevance-weighted score (Bakman et al., 2024)

$$C_{\text{rel}}(\tau, x) \;=\; \prod_{t=1}^{T} p_t^{\,w_t} \;=\; \exp\Big(\sum_{t=1}^{T} w_t \log p_t\Big). \tag{4}$$

Closest to our setting, UQAC instantiates this idea on CoT r through hard selection. It constructs an attention chain r_{attn} by backtracking from y with attention weights, optionally refining the chain by similarity to the answer. Let $\mathcal{S}$ denote the selected-token set induced by this chain (typically including answer tokens) and $S = |\mathcal{S}|$ its size. In the generic formulation above, this hard selection corresponds to $w_t = \mathbf{1}\{t \in \mathcal{S}\}$. The primary score is the unnormalized product

$$\text{UQAC}_{\text{attn}} \;=\; \prod_{t=1}^{|\mathcal{S}|} p_t\,. \tag{5}$$

Although UQAC is effective to a certain extent, the product depends jointly on these probabilities and the number of selected-token $|\mathcal{S}|$. This makes it unclear whether the probability is truly effective, and the impact of the number of selected tokens is also unclear. Therefore, the calibration source needs further investigation and we are motivated to test whether the specific selected-token probabilities are truly necessary for calibration gains and to explore better calibration signals.

3.3 Do Selected-Token Probabilities Explain UQAC's Calibration Gains?

We keep UQAC's selected-token set $\mathcal{S}$ fixed for each trajectory and construct two variants that progressively remove specific probability information. C^{S}_{mean} replaces each selected-token probability with the trajectory's mean token probability C_{mean} before taking their product, as defined in Section 3.2. C^{S}_{global} further replaces the trajectory-specific mean with C_{global}, obtained by averaging C_{mean} over all trajectories with the same model on a given dataset. This value is shared across those trajectories, so only the selected-token count S varies. Figure 2 compares these two variants, the original C_{mean}, and UQAC across Qwen2.5 family. Although C_{mean} itself has high ECE, the resulting C^{S}_{mean} achieves substantially lower ECE than UQAC. Even with a fixed constant for all trajectories from each model, C^{S}_{global} generally remains better calibrated than UQAC. Given the well calibration performance of C^{S}_{mean}, we first propose $\text{UQAC}_{\text{mean}}$ as an improved variant of UQAC, and the results in main experiments (Table 1) indeed demonstrate its advantages.

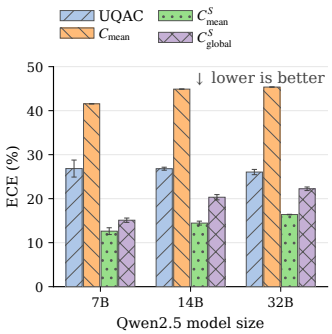

Figure 2: Using the selected count as an exponent yields the lowest ECE across model sizes.

More importantly, this pilot study shows that replacing the selected tokens' probabilities generally improves calibration over UQAC, thus the gains may not depend on the specific value of probabilities. When probabilities are replaced with a shared constant, the score depends only on the number of selected tokens, suggesting that token count may serve as a alternative calibration signal.

3.4 Motivation: Counting Divergent Tokens

To obtain such a count for calibration, we draw on inter-model disagreement as an uncertainty signal (Lakshminarayanan et al., 2017; Sun et al., 2024; Kruse et al., 2025). We hypothesize that *unreliable reasoning paths contain more tokens at which models strongly disagree*. Specifically, we use teacher forcing to measure disagreement between two models' next-token distributions at each position along the same fixed trajectory.

At reasoning position t, let $P_{\mathcal{M}}(\cdot \mid c_t)$ denote model $\mathcal{M}$'s next-token distribution under the shared prefix $c_t = (x, \tau_{<t})$. Denoting two distinct distributions as P_t and Q_t, we use Jensen–Shannon

divergence (JSD) to measure their disagreement for the token τ_t:

$$\mathrm{JSD}(P_t, Q_t) \;=\; \mathbb{H}\left(\frac{P_t + Q_t}{2}\right) - \tfrac{1}{2}\,\mathbb{H}(P_t) - \tfrac{1}{2}\,\mathbb{H}(Q_t), \tag{6}$$

where $\mathbb{H}(\cdot)$ is Shannon entropy. Models in different families disagree to different degrees, so we set the threshold θ separately for each family on a small validation set. We define a token as a *divergent token* if $\mathrm{JSD}(P_t, Q_t) > \theta$ and set of divergent token positions along a reasoning path is:

$$T_{\mathrm{div}}(\theta) \;=\; \big\{\, t \in \{1, \ldots, T\} : \mathrm{JSD}(P_t, Q_t) > \theta \,\big\}. \tag{7}$$

We use JSD for its symmetry and boundedness and see Appendix C.4 for discussion on alternative measures of token-level disagreement.

Based on the definition, we can first measure the confidence of generator $\mathcal{G}$ with an auxiliary model $\mathcal{A}$ in the *single-auxiliary* setting, which applies when $\mathcal{G}$ is a white-box model. When the target model is a black-box model and we cannot access its token distribution, we assume that *the relationship between the number of divergent tokens and reasoning-path reliability remains reasonably stable across different model pairs*, which allow us to calibrate in a *dual-auxiliary* setting by comparing two external auxiliary models from the same family, denoted as $\mathcal{A}'$ and $\mathcal{A}''$.

To verify that the number of divergent tokens $|T_{\mathrm{div}}(\theta)|$ can serve as a signal of model uncertainty, we first analyze the relationship between answer accuracy against the number of divergent tokens for different model pairs. The results are shown in Figure 1(c). In both single- and dual-auxiliary settings, accuracy decreases nearly monotonically as $|T_{\mathrm{div}}(\theta)|$ increases, supporting that the number of divergent tokens can reveal reasoning-path reliability to some extent.

4 DTC: Divergent Token Confidence

Given the divergent-token count $m = |T_{\mathrm{div}}(\theta)|$ defined in Equation (7), DTC constructs two path-level confidence estimates. The primary estimator $\mathrm{DTC}_{\mathrm{lin}}$ maps m directly to confidence, while $\mathrm{DTC}_{\mathrm{prod}}$ uses m to recalibrate the standard full-sequence confidence C_{mean} from Section 3.2.

$\mathrm{DTC}_{\mathrm{lin}}$: Count-linear confidence. Motivated by the decrease in accuracy with increasing divergent token count (Figure 1(b,c)), we define

$$\mathrm{DTC}_{\mathrm{lin}}(m) = \begin{cases} a - \dfrac{a-b}{n}\, m, & 0 \le m < n, \\ b, & m \ge n. \end{cases} \tag{8}$$

Confidence starts at a, decreases linearly with m, and reaches a floor of b at $m = n$. We use $a = 0.95$, $b = 0.05$, and $n = 10$ in all main experiments; sensitivity to n is examined in Appendix C.5. Once the divergent tokens are selected, this mapping depends only on their count.

$\mathrm{DTC}_{\mathrm{prod}}$: Trajectory-mean product confidence. Section 3.3 shows that replacing selected-token probabilities with the trajectory mean C_{mean} and taking their product reduces overconfidence. We therefore combine C_{mean} with the uncertainty-informed count m for confidence estimation:

$$\mathrm{DTC}_{\mathrm{prod}}(m) = C_{\mathrm{mean}}^{\,m+k}. \tag{9}$$

For C_{mean}, we use probabilities from the generator in white-box settings (Section 5.1) and the larger auxiliary on the same frozen trajectory in black-box settings (Section 5.2). We use $k = 4$ unless otherwise specified, avoiding a score of one solely due to a zero count and sensitivity to k is examined in Appendix C.7. When $C_{\mathrm{mean}} \in (0, 1)$, a larger m yields a lower score. Compared with $\mathrm{DTC}_{\mathrm{lin}}$, $\mathrm{DTC}_{\mathrm{prod}}$ further uses the trajectory-level probability C_{mean}, giving finer-grained confidence to trajectories that share the same m. The uncertainty-informed count m can also be combined with other overconfident scores to improve calibration (Section 6).

5 Experiments

5.1 White-box Setting

Datasets and Models. To evaluate DTC on tasks with long CoT path, we use four widely used mathematical benchmarks spanning a range of difficulty: MATH-500 (Hendrycks et al., 2021; Light-

Table 1: Uncertainty quantification performance (ECE) in white-box settings. In each row, the best result is in **bold** and the second-best one is underlined (excluding PRM). †: UQAC variant by ours.

Reasoning Models	Acc	C_{NSL}	C_{mean}	Entropy Conf.	BaseCal	UQAC		Verb.	DTC (Ours)		PRM (ref.)
						attn	mean†		prod	lin	
MATH-500											
Qwen2.5-7B	76.1	43.6	45.3	32.8	43.7	26.8	**12.7**	44.2	20.0	13.4	9.8
Qwen2.5-14B	80.0	43.0	44.9	38.9	43.0	26.8	14.2	34.7	16.2	**9.9**	8.8
Qwen2.5-32B	82.4	43.7	45.4	39.9	43.8	26.1	16.2	41.6	16.9	**6.6**	8.4
Qwen3-8B	83.9	38.1	41.5	31.8	36.7	37.1	37.8	47.8	**7.9**	15.6	7.4
Qwen3-14B	86.5	36.8	40.6	29.9	36.2	27.1	13.2	42.5	**5.5**	11.6	6.2
Qwen3-32B	83.7	36.1	40.2	29.4	—	38.8	38.6	37.0	**6.1**	11.0	6.3
Gemma3-12B	84.9	41.5	44.0	37.3	34.8	33.9	29.3	24.3	15.2	**13.9**	7.2
Gemma3-27B	89.2	42.3	44.5	38.4	36.0	16.0	**7.0**	33.4	15.0	7.1	8.4
AMC23											
Qwen2.5-7B	53.6	43.6	45.3	31.7	44.0	24.7	8.2	42.5	15.7	**6.7**	6.0
Qwen2.5-14B	61.2	42.7	44.7	38.5	43.1	25.3	19.7	35.1	9.7	**9.5**	6.0
Qwen2.5-32B	66.4	43.3	45.1	39.5	44.0	24.9	22.7	38.3	**9.8**	10.8	8.9
Qwen3-8B	68.6	36.5	40.5	29.4	36.8	49.4	38.5	48.7	**6.8**	9.3	11.4
Qwen3-14B	74.1	34.6	39.2	26.9	35.9	27.2	**10.5**	35.3	12.3	11.8	12.0
Qwen3-32B	67.5	33.7	38.5	25.4	—	44.0	37.7	32.3	15.2	**7.5**	11.6
Gemma3-12B	66.8	40.8	43.6	36.4	35.3	34.5	24.0	25.9	12.7	**10.9**	10.1
Gemma3-27B	76.9	41.8	44.3	37.9	36.9	19.9	12.3	26.4	12.4	**12.0**	13.1
AIME24											
Qwen2.5-7B	12.6	42.8	44.7	30.3	43.3	29.5	19.0	39.5	14.7	**11.5**	22.7
Qwen2.5-14B	13.8	41.8	44.0	37.5	42.3	25.1	15.6	31.4	**13.1**	16.9	23.0
Qwen2.5-32B	16.9	42.4	44.4	37.9	43.5	23.9	16.7	34.1	**10.6**	20.4	23.4
Qwen3-8B	27.7	34.9	39.3	26.5	36.4	40.3	42.9	45.8	**14.2**	14.9	29.2
Qwen3-14B	27.1	33.2	38.1	24.4	35.7	31.4	**16.2**	44.1	20.6	18.1	29.7
Qwen3-32B	28.3	32.0	37.4	23.7	—	44.2	37.6	25.7	21.6	**16.4**	28.8
Gemma3-12B	23.9	39.8	42.8	34.7	34.6	33.2	31.8	23.0	**9.7**	12.6	29.2
Gemma3-27B	29.0	40.7	43.5	36.2	36.1	29.6	29.4	21.1	**8.1**	24.3	30.0
AIME25											
Qwen2.5-7B	9.1	42.4	44.4	31.0	43.1	26.6	16.2	37.3	**11.2**	20.5	25.2
Qwen2.5-14B	14.7	42.4	44.4	37.8	42.9	27.5	15.7	31.6	**10.8**	25.4	24.4
Qwen2.5-32B	12.2	42.7	44.7	38.3	43.8	30.6	18.4	33.0	**8.7**	28.3	27.7
Qwen3-8B	22.5	34.2	38.8	25.9	36.0	37.6	43.4	46.9	12.3	**9.5**	29.6
Qwen3-14B	27.5	32.6	37.8	24.0	35.3	30.8	19.7	53.0	17.1	**5.3**	30.7
Qwen3-32B	23.7	31.7	37.2	22.5	—	36.3	37.8	25.6	19.1	**6.9**	30.8
Gemma3-12B	18.8	40.3	43.2	35.5	34.5	32.1	29.4	15.9	16.7	**7.8**	27.2
Gemma3-27B	25.3	40.6	43.4	35.9	35.4	25.1	22.7	23.1	15.6	**11.1**	26.7
Average	48.0	39.3	42.4	32.7	39.0	30.8	23.6	35.0	13.2	**13.0**	18.1

man et al., 2024) and the contest sets AMC23, AIME24, and AIME25 (Balunovic et al., 2025). We evaluate eight LLMs from three families: Qwen2.5-{7,14,32}B-Instruct (Qwen Team, 2024), Qwen3-{8,14,32}B (Yang et al., 2025), and Gemma3-{12,27}B-IT (Gemma Team, 2025).

Baselines. Given the limited prior work on calibrated reasoning uncertainty, we compare DTC against three classes of uncertainty estimators. (1) *Standard full-sequence confidence*: C_{NSL} uses length-normalized sequence likelihood (Equation (3)); C_{mean} uses mean token probability (Section 3.2); and **Entropy Confidence** that uses length-normalized predictive entropy confidence (Li et al., 2025). (2) *Refinements of full-sequence confidence*: **BaseCal** (Tan et al., 2026) averages response-token probabilities under a paired base model; $\text{UQAC}_{\text{attn}}$ (Li et al., 2025) further select attention-related tokens; and our variant $\text{UQAC}_{\text{mean}}$ which replaces each selected-token probability with the C_{mean} before multiplying over the selected set. BaseCal is omitted for Qwen3-32B because no public base checkpoint is available. (3) *Verbalized estimation*: Xiong et al. (2024) propose **Verbalized confidence** through additional prompts after the response elicits the model confidence. We also include **PRM** (Skywork-o1-Open-PRM-Qwen-2.5-7B; He et al., 2024) as a supervised reference, averaging rewards over newline-delimited reasoning steps into path-level confidence.

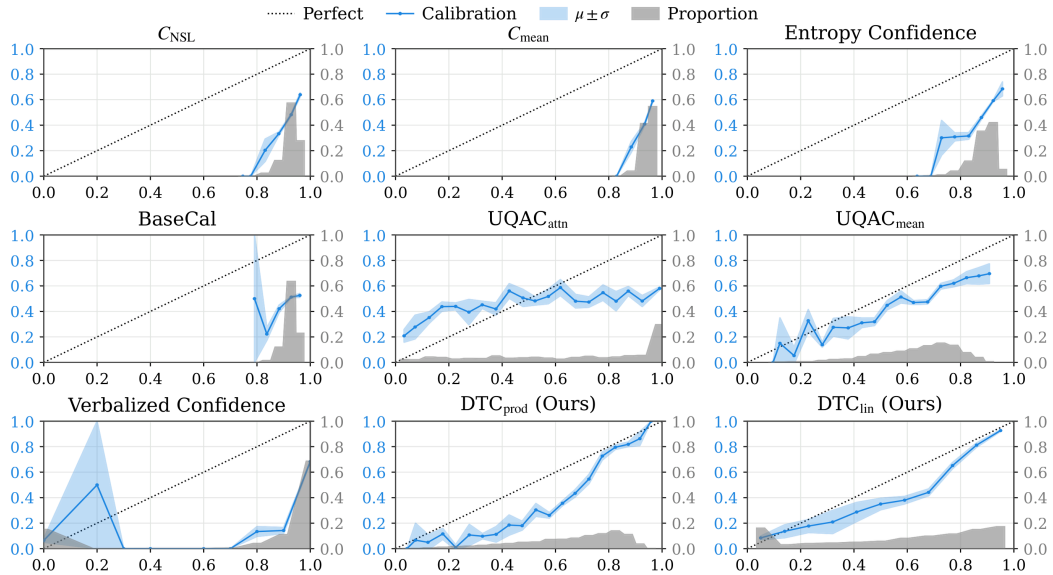

Figure 3: Calibration plots and probability histogram for Qwen2.5-14B on MATH-500. The x-axis shows mean confidence within 20 probability bins. The calibration curve (blue line with $\mu \pm \sigma$) displays actual accuracy per bin, while the gray shadow represents the probability proportion.

Implementations. We use the officially recommended decoding parameters for each LLM, as detailed in Appendix A.2. For each model family, we use its smallest model and 100 problems drawn at random from the MATH training set (Hendrycks et al., 2021) to set that family's disagreement threshold θ. We report DTC_{prod} and DTC_{lin}, with $\theta = 0.50$ for Qwen2.5, $\theta = 0.60$ for Qwen3, and $\theta = 0.85$ for Gemma3. For DTC_{prod}, token probabilities come from the generator $\mathcal{G}$ (Equation (9)). Unless noted, the auxiliary is the smaller same-family instruct model: Qwen2.5-1.5B-Instruct, Qwen3-1.7B, or Gemma3-4B-IT. Other auxiliary sizes are examined in Section 6. Following Li et al. (2025), we evaluate ECE and AUROC by repeated subsampling, emphasizing ECE and the calibration plots. AUROC is the area under the ROC curve and measures ranking. Tables and figures use AUC for AUROC. Further details are in Appendix A.3.

Main Results. Table 1 and Table 5 (Appendix B.1) report the ECE and AUROC results, respectively. Figure 3 visualizes calibration curves on MATH-500 with Qwen2.5-14B. We find that:

Full-sequence confidence retains ranking information but remains overconfident. The three standard methods (C_{NSL}, C_{mean} and Entropy Confidence) achieve an high average AUROC of 72.7–73.8, yet their ECEs reach 32.7–42.4. Their calibration curves lie below the diagonal, with scores concentrated near the high-confidence.

Refinements improve calibration unevenly and weaken ranking. BaseCal offers limited calibration gains, whereas the UQAC variants reduce ECE more substantially. UQAC variant $\text{UQAC}_{\text{mean}}$ by ours achieves both the lowest average ECE (23.6) and highest AUROC (66.7) among these refinements, but its ranking still trails the standard methods.

DTC improves both calibration and average ranking. DTC_{lin} and DTC_{prod} reduce average ECE to 13.0 and 13.2, with calibration curves closer to the diagonal, while attaining AUROC 75.3 and 80.8. Thus, count alone supports effective calibration; retaining path-mean probability further improves average ranking at similar ECE. Relative to $\text{UQAC}_{\text{mean}}$, DTC_{prod} lowers average ECE by 10.4, suggesting that selecting divergent tokens captures uncertainty more effectively and improves the calibration of C_{mean}. In harder datasets, PRM has higher ECE, while DTC is better calibrated.

5.2 Black-box Setting

Datasets and Models. To better match a black-box evaluation setting, we use more challenging contest benchmarks than in the white-box experiments, namely AIME24, AIME25, HMMT25, and

Table 2: Uncertainty quantification performance with DeepSeek-V3.2 in black-box settings.

Method	AIME24			AIME25			HMMT25			HMMT26		
	Acc↑	ECE↓	AUC↑	Acc↑	ECE↓	AUC↑	Acc↑	ECE↓	AUC↑	Acc↑	ECE↓	AUC↑
Default CoT	72.2	–	–	60.9	–	–	46.1	–	–	50.9	–	–
→ + PRM	–	39.0	**88.2**	–	39.5	80.4	–	41.1	60.4	–	38.1	75.1
→ + $\mathrm{DTC}_{\mathrm{lin}}$	–	**11.2**	74.6	–	**9.4**	78.7	–	**12.7**	67.9	–	**12.5**	72.7
→ + $\mathrm{DTC}_{\mathrm{prod}}$	–	24.6	79.4	–	23.9	**82.1**	–	29.1	**75.0**	–	26.4	**76.5**
Verb. Conf.	67.4	40.8	**85.5**	55.3	39.6	**86.8**	38.3	40.3	**78.2**	45.0	40.1	**84.3**
→ + $\mathrm{DTC}_{\mathrm{lin}}$	–	**13.5**	67.1	–	**9.2**	73.5	–	**17.5**	59.5	–	**14.5**	68.9
→ + $\mathrm{DTC}_{\mathrm{prod}}$	–	25.9	73.5	–	25.4	78.3	–	30.9	67.0	–	26.1	73.7
Verb. TopK	64.6	32.0	**88.4**	52.8	32.6	**85.7**	27.1	31.6	**79.3**	35.1	32.0	**83.6**
→ + $\mathrm{DTC}_{\mathrm{lin}}$	–	**14.2**	66.0	–	**14.1**	66.9	–	**18.8**	56.7	–	**18.2**	60.9
→ + $\mathrm{DTC}_{\mathrm{prod}}$	–	32.9	71.4	–	31.4	71.2	–	36.1	61.5	–	31.1	64.6
Verb. PD	68.3	32.6	**86.2**	57.1	33.2	**84.9**	33.1	31.5	**74.3**	37.9	32.3	**83.1**
→ + $\mathrm{DTC}_{\mathrm{lin}}$	–	**15.1**	65.4	–	**13.5**	69.1	–	**19.2**	57.8	–	**14.6**	66.0
→ + $\mathrm{DTC}_{\mathrm{prod}}$	–	35.3	72.6	–	32.6	74.1	–	36.2	63.0	–	31.8	69.3

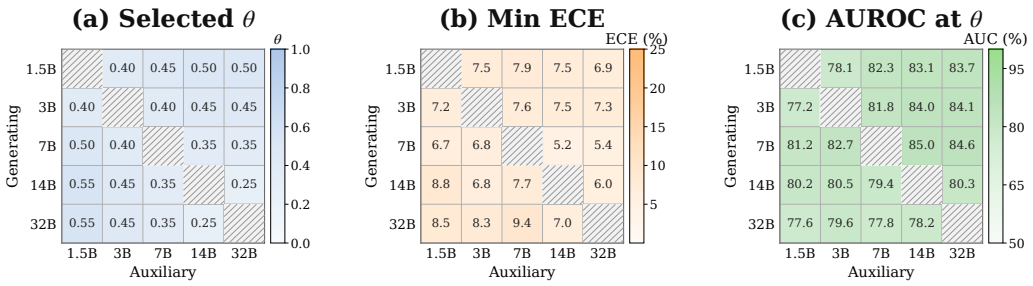

Figure 4: Sensitivity of $\mathrm{DTC}_{\mathrm{lin}}$ to θ and auxiliary size, single-auxiliary Qwen2.5 on AMC23. We search θ in steps of 0.05 and report the θ (a) with the lowest ECE (b) and its AUROC (c).

HMMT26 (Balunovic et al., 2025), and switch to stronger generators: Qwen3-4B-Instruct-2507, Qwen3-30B-A3B-Instruct-2507 (Yang et al., 2025), and DeepSeek-V3.2 (DeepSeek-AI, 2025). We treat each generator as a black box, using its output trajectories without accessing its logits for scoring. More details on datasets and models are shown in Appendix A.1.

Baselines. Because generator token probabilities are unavailable, we use Verbalized estimation methods based on CoT prompts as the main confidence baselines (Wang & Zhang, 2026). We consider three Verbalized estimation methods: **Verbalized Confidence** (Verb. Conf.), **Verbalized TopK** (Verb. TopK) (Tian et al., 2023), and **Verbalized Probability Distribution** (Verb. PD) (Wang et al., 2025). As described in Section 2, these Verbalized estimation methods change the CoT trajectory, so besides applying our method on **Default CoT**, we also apply it to the trajectories generated by these prompts for a fair comparison.

Evaluation and Implementation. Metrics, evaluation, and decoding follow the white-box setting; details appear in Appendix A.2 and A.3. All three generators use Qwen2.5-7B-Instruct as $\mathcal{A}'$ and Qwen2.5-1.5B-Instruct as $\mathcal{A}''$ (Qwen Team, 2024), with θ=0.70.

Main Results. Tables 2, 6 and 7 report accuracy, ECE, and AUROC across three generators and four mathematics benchmarks, respectively. Blue arrows (→) in the method column denote rescoring the same trajectories, leaving answers and accuracy unchanged. PRM additionally scores Default CoT trajectories as a reference. **Verbalized estimation is overconfident and sometimes reduces answer accuracy.** Consistent with the accuracy and calibration limitations observed in mathematical reasoning settings by Wang et al. (2025), we find that Verbalized estimation can reduce answer accuracy while producing overconfident scores. For DeepSeek-V3.2, the three Verbalized estimation methods reduce mean accuracy from 57.5 under Default CoT to 44.9–51.5, while their verbalized scores have mean ECE values of 32.1–40.2.

DTC achieves low ECE on Default CoT and improves calibration over verbalized scores on the same trajectories. In contrast, our method assesses the reliability of Default CoT trajectories without changing the generated answers or reducing model accuracy. On Default CoT, $\mathrm{DTC}_{\mathrm{lin}}$ achieves mean ECE values of 11.5–12.7 across the three generators. When applied to trajectories generated under Verb. Conf., Verb. TopK, and Verb. PD, it also yields lower ECE than the original verbalized scores for every model–dataset pair. On DeepSeek-V3.2, its mean ECE is 13.7–16.3, compared with 32.1–40.2 for the original verbalized scores. $\mathrm{DTC}_{\mathrm{prod}}$ achieves higher mean AUROC than $\mathrm{DTC}_{\mathrm{lin}}$, but also has higher mean ECE. Despite these calibration gains, both estimators often have lower AUROC than the verbalized scores on trajectories from these three Verbalized estimation methods.

6 Analyses and Discussions

Disagreement Threshold and auxiliary size. We study how the choice of auxiliary model and disagreement threshold θ affects $\mathrm{DTC}_{\mathrm{lin}}$. For each pair of generator and auxiliary, we search θ in steps of 0.05 and report the θ with the lowest ECE (Figure 4). The lowest ECE is similar across auxiliary sizes. At these values, a larger auxiliary usually has a higher AUROC. When the auxiliary is fixed at 1.5B, the θ with the lowest ECE stays near 0.5 (Figure 5). For efficiency and a simpler method, each model family uses one small auxiliary from the same family and one fixed θ. The corresponding dual-auxiliary analyses are given in Appendix C.1; the overall conclusions are similar to those in the single-auxiliary setting.

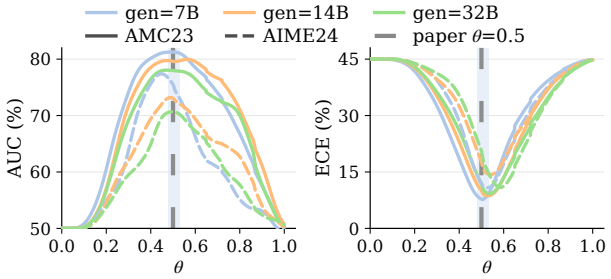

Figure 5: $\mathrm{DTC}_{\mathrm{lin}}$ sensitivity to θ with auxiliary fixed at 1.5B and generating models $\in \{7, 14, 32\}$B (solid: AMC23; dashed: AIME24). The shaded band marks the white-box operating point θ=0.50.

Combining verbalized scores with DTC. Verbalized estimation often yields overconfident scores (high ECE) despite competitive ranking (high AUROC). We examine whether combining the divergent-token count with verbalized scores can reduce the overconfidence of Verbalized estimation, as $\mathrm{DTC}_{\mathrm{prod}}$ does for C_{mean}. We follow the way $\mathrm{DTC}_{\mathrm{prod}}$ combines C_{mean} with the uncertainty-informed count m, and combine verbalized scores with DTC to get p_{verb}^{m+k}. The auxiliary and k follow the black-box $\mathrm{DTC}_{\mathrm{prod}}$ setting. We report the result as w/ DTC in Tables 3, 8 and 9. Across these tables, DTC lowers ECE in every case, often by more than 20 percentage points. AUROC changes little and sometimes improves. On DeepSeek-V3.2, ECE decreases by 12.3 to 35.9 percentage points, and AUROC changes by -0.4 to $+2.0$ percentage points. Combining verbalized confidence with DTC thus gives a better tradeoff between ECE and AUROC.

Table 3: Verbalized vs. w/ DTC on DeepSeek-V3.2.

Method	AIME24		AIME25		HMMT25		HMMT26	
	ECE↓	AUC↑	ECE↓	AUC↑	ECE↓	AUC↑	ECE↓	AUC↑
Verb. Conf.	40.8	85.5	39.6	86.8	40.3	78.2	40.1	84.3
w/ DTC	6.3	85.8	3.7	87.7	10.5	77.8	6.2	84.2
Verb. TopK	32.0	88.4	32.6	85.7	31.6	79.3	32.0	83.6
w/ DTC	17.0	90.3	14.6	87.3	18.8	79.1	15.7	83.6
Verb. PD	32.6	86.2	33.2	84.9	31.5	74.3	32.3	83.1
w/ DTC	10.2	87.7	10.9	86.9	19.2	76.0	13.9	84.0

7 Conclusion

We study the reliability of a reasoning path. The UQAC pilot indicates that much of the calibration gain comes from how many tokens are selected. DTC counts tokens on which two models disagree strongly about the next token. $\mathrm{DTC}_{\mathrm{lin}}$ takes this count as the confidence. $\mathrm{DTC}_{\mathrm{prod}}$ recalibrates the trajectory's mean token probability with the same count, and a larger count gives a lower score. On the math benchmarks, both improve calibration over sequence likelihood, verbalized confidence, and UQAC, whether the generator logits are available or not. Applied to a verbalized score, the count also reduces overconfidence.

AI USE STATEMENT

We used generative AI tools to implement evaluation, plotting, and table-generation code, to assist with translation, and to draft and edit manuscript text, figures, and related-work notes. We have not used generative AI tools to generate synthetic datasets, to formulate or prove mathematical claims, or to interpret experimental results, and the remaining required-disclosure tasks are not applicable. Additionally, we used generative AI tools to search and format references and to improve readability. We reviewed all AI-assisted text, citations, and code; numbers reported in the paper come from our evaluation pipeline and were checked by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Datasets, models, sampling hyperparameters, operating thresholds, and baseline protocols are specified in Section 5 and Appendix A. Estimator definitions appear in Section 4; additional ablations and black-box settings are in Appendix C and Appendix C.5.

REFERENCES

Yavuz Faruk Bakman, Duygu Nur Yaldiz, Baturalp Buyukates, Chenyang Tao, Dimitrios Dimitriadis, and Salman Avestimehr. MARS: Meaning-aware response scoring for uncertainty estimation in generative LLMs. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7752–7767, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.419. URL https://aclanthology.org/2024.acl-long.419/.

Mislav Balunovic, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin Vechev. Matharena: Evaluating llms on uncontaminated math competitions. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Datasets and Benchmarks Track, pp. 22851–22888. Curran Associates, Inc., 2025. doi: 10.52202/085713-0679. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/1d27c01ebd3e3aebe226b44fc970d803-Paper-Datasets_and_Benchmarks_Track.pdf.

Guangsheng Bao, Hongbo Zhang, Cunxiang Wang, Linyi Yang, and Yue Zhang. How likely do LLMs with CoT mimic human reasoning? In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 7831–7850, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.524/.

Jan Clusmann, Fiona R. Kolbinger, Hannah Sophie Muti, Zunamys I. Carrero, Jan-Niklas Eckardt, Narmin Ghaffari Laleh, Chiara Maria Lavinia Löffler, Sophie-Caroline Schwarzkopf, Michaela Unger, Gregory P. Veldhuizen, Sophia J. Wagner, and Jakob Nikolas Kather. The future landscape of large language models in medicine. *Communications Medicine*, 3(1):141, 2023. doi: 10.1038/s43856-023-00370-1. URL https://doi.org/10.1038/s43856-023-00370-1.

DeepSeek-AI. DeepSeek-V3.2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*, 2025. URL https://arxiv.org/abs/2512.02556.

Siddartha Devic, Tejas Srinivasan, Jesse Thomason, Willie Neiswanger, and Vatsal Sharan. From calibration to collaboration: LLM uncertainty quantification should be more human-centered. *arXiv preprint arXiv:2506.07461*, 2025. URL https://arxiv.org/abs/2506.07461.

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5050–5063, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.276. URL https://aclanthology.org/2024.acl-long.276/.

Yichao Fu, Xuewei Wang, Hao Zhang, Yuandong Tian, and Jiawei Zhao. Deep think with confidence. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 94355–94377, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/98257285340854262185500e59bc0f28-Paper-Conference.pdf.

Aryo Pradipta Gema, Alexander Hägele, Runjin Chen, Andy Arditi, Jacob Goldman-Wetzler, Kit Fraser-Taliente, Henry Sleight, Linda Petrini, Julian Michael, Beatrice Alex, Pasquale Minervini, Yanda Chen, Joe Benton, and Ethan Perez. Inverse Scaling in Test-Time Compute. *Transactions on Machine Learning Research*, 2025. URL https://mlanthology.org/tmlr/2025/gema2025tmlr-inverse/.

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. URL https://arxiv.org/abs/2503.19786.

Soumya Suvra Ghosal, Souradip Chakraborty, Avinash Reddy, Yifu Lu, Mengdi Wang, Dinesh Manocha, Furong Huang, Mohammad Ghavamzadeh, and Amrit Singh Bedi. Does thinking more always help? mirage of test-time scaling in reasoning models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 191359–191386. Curran Associates, Inc., 2025. doi: 10.52202/085713-5747. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/fc067ac218430c409d6f65403328f740-Paper-Conference.pdf.

Matt Gorbett and Suman Jana. Cross-model disagreement as a label-free correctness signal. *arXiv preprint arXiv:2603.25450*, 2026. URL https://arxiv.org/abs/2603.25450.

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 - Volume 70*, ICML'17, pp. 1321–1330. JMLR.org, 2017.

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, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, 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, Mingxu Zhou, 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, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, 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, 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 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025. doi: 10.1038/s41586-025-09422-z. URL https://doi.org/10.1038/s41586-025-09422-z.

Jujie He, Tianwen Wei, Rui Yan, Jiacai Liu, Chaojie Wang, Yimeng Gan, Shiwen Tu, Chris Yuhao Liu, Liang Zeng, Xiaokun Wang, Boyang Wang, Yongcong Li, Fuxiang Zhang, Jiacheng Xu, Bo An, Yang Liu, and Yahui Zhou. Skywork-o1 open series, November 2024. URL https://doi.org/10.5281/zenodo.16998085.

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, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/be83ab3ecd0db773eb2dc1b0a17836a1-Paper-round2.pdf.

Maya Kruse, Majid Afshar, Saksham Khatwani, Anoop Mayampurath, Guanhua Chen, and Yanjun Gao. Simple yet effective: An information-theoretic approach to multi-LLM uncertainty quantification. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 30493–30504, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1551. URL https://aclanthology.org/2025.emnlp-main.1551/.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=VD-AYtP0dve.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf.

Yinghao Li, Rushi Qiang, Lama Moukheiber, and Chao Zhang. Language model uncertainty quantification with attention chain. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=QTrW2HWNXe.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/aca97732e30bcf1303bc22ac3924fd16-Paper-Conference.pdf.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Contextualized sequence likelihood: Enhanced confidence scores for natural language generation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 10351–10368, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.578. URL https://aclanthology.org/2024.emnlp-main.578/.

Xiaoou Liu, Tiejin Chen, Longchao Da, Chacha Chen, Zhen Lin, and Hua Wei. Uncertainty quantification and confidence calibration in large language models: A survey. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, pp. 6107–6117, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400714542. doi: 10.1145/3711896.3736569. URL https://doi.org/10.1145/3711896.3736569.

Nishanth Madhusudhan, Sathwik Tejaswi Madhusudhan, Vikas Yadav, and Masoud Hashemi. Do LLMs know when to NOT answer? investigating abstention abilities of large language models. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 9329–9345, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.627/.

Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=jN5y-zb5Q7m.

OpenAI. OpenAI o1 system card. *arXiv preprint arXiv:2412.16720*, 2024. URL https://arxiv.org/abs/2412.16720.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. Llms know more than they show: On the intrinsic representation of llm hallucinations. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.),

International Conference on Learning Representations, volume 2025, pp. 66880–66913, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/a712d461e57201efe35d429a6f1731c1-Paper-Conference.pdf.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Guangzhi Sun, Potsawee Manakul, Adian Liusie, Kunat Pipatanakul, Chao Zhang, Philip C. Woodland, and Mark J. F. Gales. Crosscheckgpt: Universal hallucination ranking for multimodal foundation models. *CoRR*, abs/2405.13684, 2024. URL https://doi.org/10.48550/arXiv.2405.13684.

Hexiang Tan, Wanli Yang, Junwei Zhang, Xin Chen, Rui Tang, Du Su, Jingang Wang, Yuanzhuo Wang, Fei Sun, and Xueqi Cheng. BaseCal: Unsupervised confidence calibration via base model signals. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5172–5187, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.234. URL https://aclanthology.org/2026.acl-long.234/.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5433–5442, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.330. URL https://aclanthology.org/2023.emnlp-main.330/.

Ante Wang, Weizhi Ma, and Yang Liu. Let the model distribute its doubt: Confidence estimation through verbalized probability distribution. *arXiv preprint arXiv:2511.14275*, 2025. URL https://arxiv.org/abs/2511.14275.

Jiayi Wang and Xu-Yao Zhang. A systematic evaluation of black-box uncertainty estimation methods for large language models. *arXiv preprint arXiv:2606.19868*, 2026. URL https://arxiv.org/abs/2606.19868.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837. Curran Associates, Inc., 2022. doi: 10.52202/068431-1800. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 23650–23678, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/6733cf15e10e2cd1d59af033c3bb8507-Paper-Conference.pdf.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Boxuan Zhang and Ruqi Zhang. CoT-UQ: Improving response-wise uncertainty quantification in LLMs with chain-of-thought. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 26114–26133, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1339. URL https://aclanthology.org/2025.findings-acl.1339/.

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. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 52910–52942, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/56b06e61ddd2600fe86cca96f53869f6-Paper-Conference.pdf.

(Andrew) Zhanke Zhou, Zhaocheng Zhu, Xuan Li, Mikhail Galkin, Xiao Feng, Sanmi Koyejo, Jian Tang, and Bo Han. Landscape of thoughts: Visualizing the reasoning process of large language models. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 74411–74466, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/79201c5a56d6f39520fb8a06b1ec415c-Paper-Conference.pdf.

A Experimental Implementation Details

This appendix records the experimental setup for Sections 5.1 and 5.2: benchmarks and models, sampling, answer evaluation, confidence computation, and baselines.

A.1 Datasets and Models

We use six math benchmarks. White-box experiments use MATH-500, AMC23, AIME24, and AIME25; black-box experiments use AIME24, AIME25, HMMT25, and HMMT26. AIME24 and AIME25 appear in both. Table 4 lists the generators, auxiliary models, and sampling budgets.

Table 4: Benchmarks, models, and sample counts per question. The three black-box generators share one auxiliary pair.

	White-box	Black-box
Benchmarks	MATH-500 (Lightman et al., 2024) AMC23 (Balunovic et al., 2025) AIME24 (Balunovic et al., 2025) AIME25 (Balunovic et al., 2025)	AIME24 (Balunovic et al., 2025) AIME25 (Balunovic et al., 2025) HMMT25 (Balunovic et al., 2025) HMMT26 (Balunovic et al., 2025)
Generators	Qwen2.5-7B-Instruct (Qwen Team, 2024) Qwen2.5-14B-Instruct (Qwen Team, 2024) Qwen2.5-32B-Instruct (Qwen Team, 2024) Qwen3-8B (Yang et al., 2025) Qwen3-14B (Yang et al., 2025) Qwen3-32B (Yang et al., 2025) Gemma3-12B-IT (Gemma Team, 2025) Gemma3-27B-IT (Gemma Team, 2025)	DeepSeek-V3.2 (DeepSeek-AI, 2025) Qwen3-30B-A3B-Instruct-2507 (Yang et al., 2025) Qwen3-4B-Instruct-2507 (Yang et al., 2025)
Auxiliaries	Qwen2.5-1.5B-Instruct (Qwen Team, 2024) Qwen3-1.7B (Yang et al., 2025) Gemma3-4B-IT (Gemma Team, 2025)	Qwen2.5-7B-Instruct (Qwen Team, 2024) Qwen2.5-1.5B-Instruct (Qwen Team, 2024)
Sample number per question	8 on MATH-500 64 on AMC23 64 on AIME24 64 on AIME25	32 for DeepSeek-V3.2 64 for Qwen3-30B-A3B-Instruct-2507 64 for Qwen3-4B-Instruct-2507

White-box generators are Qwen2.5-{7B,14B,32B}-Instruct, Qwen3-{8B,14B,32B}, and Gemma3-{12B,27B}-IT, paired with the smaller same-family auxiliaries Qwen2.5-1.5B-Instruct, Qwen3-1.7B, and Gemma3-4B-IT. Black-box generators are DeepSeek-V3.2, Qwen3-30B-A3B-Instruct-2507, and Qwen3-4B-Instruct-2507. In the main black-box setting, all three are scored with the external pair Qwen2.5-7B-Instruct and Qwen2.5-1.5B-Instruct.

A.2 Sampling and Prompting

We use zero-shot prompts. We ask the model to solve the problem step by step and put the final answer in a boxed expression:

Reasoning Prompt

Solve the problem step by step. Put the final answer in a boxed expression.

For the black-box protocols in Section 5.2, we add each protocol's confidence instruction at generation time. We then score these trajectories and do not generate new ones.

We decode each generator with its own recommended setting. We set the maximum sequence length to 8192 tokens for all white-box generators and to 16384 tokens for all black-box generators. Under vLLM, we use temperature 0.6 and top-$p = 0.95$ for Qwen2.5, non-2507 Qwen3, and Gemma3-IT. For Qwen3 Instruct-2507, we use temperature 0.7, top-$p = 0.8$, top-$k = 20$, and presence penalty 1.0. We sample DeepSeek-V3.2 from its API at temperature 1.0 and top-$p = 0.95$. We draw 8 trajectories per MATH-500 problem and 64 per contest problem for vLLM generators, and 32 per problem for DeepSeek-V3.2.

A.3 Answer Extraction and Evaluation

Following Li et al. (2025), we separate answer generation from uncertainty quantification. Once a response is generated, we extract the final answer and check whether it is correct. If no answer is extracted, we exclude that instance from the UQ evaluation. We then compute confidence on the same trajectory with teacher forcing. This step does not generate a new answer.

Accuracy counts every sampled response, including truncated ones. The denominator is the number of samples, and the numerator is the number labeled `correct=true`.

We then subsample to balance the number of correct and incorrect predictions, again following Li et al. (2025). If there are fewer correct predictions than incorrect ones, we randomly sample a matching number of incorrect predictions. If both groups exceed 500 instances, we randomly select 500 from each group. This step does not change the confidence scores, but it does change AUROC and ECE. We repeat the subsampling five times and report the mean and standard deviation. Accuracy is left unchanged.

ECE uses 20 equal-width bins; lower is better. AUROC is the secondary ranking metric and is unchanged under a monotone rescaling of the confidence score. Unless noted, every compared score is computed on the trajectories used for the correctness labels.

A.4 Confidence Computation

On each frozen trajectory we run teacher forcing on the shared reasoning prefixes, compute the next-token JSD, and count positions whose divergence exceeds θ (Equation (7)). In the white-box setting the two distributions are the generator and a smaller same-family auxiliary. In the black-box setting they are Qwen2.5-7B-Instruct and Qwen2.5-1.5B-Instruct, so the generator's logits are not used.

Operating points are $\theta = 0.50$ for Qwen2.5, $\theta = 0.60$ for Qwen3, $\theta = 0.85$ for Gemma3, and $\theta = 0.70$ in the black-box setting. The count map uses $a = 0.95$, $b = 0.05$, and $n = 10$; the product estimator uses $k = 4$ unless noted (Section 4). For $\mathrm{DTC}_{\mathrm{prod}}$, C_{mean} comes from the generator in white-box runs and from the larger auxiliary on the same trajectory in black-box runs. Sweeps over the threshold and the saturation length are in Appendix C.5.

A.5 Baselines

We describe the white-box baselines in the three groups used in Table 1. When a score needs token probabilities, we use the generator $\mathcal{G}$. We omit BaseCal for Qwen3-32B because no public base checkpoint is available.

Standard full-sequence confidence. These scores pool the generator's token probabilities, or its full predictive distribution, over the whole trajectory τ.

Length-normalized sequence likelihood. The product of token probabilities shrinks as the trajectory grows, so we take its geometric mean (Equation (3); Malinin & Gales, 2021). With $p_t = P_{\mathcal{G}}(\tau_t \mid x, \tau_{<t})$ and length T,

$$C_{\mathrm{NSL}} = \Big(\prod_{t=1}^{T} p_t\Big)^{1/T}. \tag{10}$$

Mean token probability. The arithmetic mean also removes the length effect, and a single low-probability token moves it less than a product (Orgad et al., 2025):

$$C_{\mathrm{mean}} = \frac{1}{T}\sum_{t=1}^{T} p_t. \tag{11}$$

Entropy confidence. Following the UQAC implementation (Li et al., 2025), at each position we keep the five largest next-token probabilities from the softmax and do not renormalize them. Predictive entropy is the sum of these per-position entropies. We divide that sum by the number of

positions and subtract the result from 1:

$$H_t \;=\; -\sum_{k=1}^{5} p_{t,(k)} \log p_{t,(k)}, \qquad 1 - \bar{H}(\tau) \;=\; 1 - \frac{1}{T}\sum_{t=1}^{T} H_t, \tag{12}$$

where $p_{t,(k)}$ is the k-th largest next-token probability at position t.

Refinements of full-sequence confidence. These scores change which probabilities enter the aggregate: either by rereading the response with a base model, or by keeping only tokens tied to the answer.

BaseCal. BaseCal scores a post-trained generator with its paired base model, which is often better calibrated (Tan et al., 2026). BaseCal-ReEval passes the same trajectory through the base model and averages the probabilities of the generated tokens:

$$C_{\mathrm{ReEval}}(\tau, x) \;=\; \frac{1}{T}\sum_{t=1}^{T} P_{\mathrm{base}}(\tau_t \mid x, \tau_{<t}). \tag{13}$$

BaseCal-Proj instead trains a one-layer linear map from the generator's final-layer hidden states into the base model's hidden states, then reads token probabilities from the base output layer. On our out-of-domain data this variant calibrates poorly, so we report only BaseCal-ReEval.

$\mathrm{UQAC}_{\mathrm{attn}}$. Many tokens in a long trajectory say little about the answer. UQAC backtracks from the answer along attention and multiplies only the selected token probabilities (Li et al., 2025). With selected set $\mathcal{S}$,

$$\mathrm{UQAC}_{\mathrm{attn}} \;=\; \prod_{t \in \mathcal{S}} p_t. \tag{14}$$

$\mathrm{UQAC}_{\mathrm{mean}}$. The product in Equation 14 still depends on each selected probability. We replace every selected probability with the trajectory mean C_{mean}, so the score moves mainly with the number of selected tokens:

$$\mathrm{UQAC}_{\mathrm{mean}} \;=\; C_{\mathrm{mean}}^{|\mathcal{S}|}. \tag{15}$$

Verbalized estimation. Following Li et al. (2025), we obtain the score by an additional prompt after the response (Xiong et al., 2024). The prompt asks the model to rate its previous answer from 0 to 10 and place the integer in `\boxed{}`. We divide that integer by 10, which gives a confidence in $[0, 1]$.

Process-reward reference. We also report Skywork-o1-Open-PRM-Qwen-2.5-7B as a supervised reference (He et al., 2024). We split the response at newlines. For each step we apply a sigmoid to the flagged token value and then average the step rewards. PRM is left out of the best and second-best marks in the white-box table.

Black-box baselines. The generator's token probabilities are unavailable, so we use three verbalized protocols. The prompts follow Wang et al. (2025) and are lightly adapted for mathematical answers (Figures 6 to 8). An answer is a mathematical value or a short exact expression, and if the question specifies a report form such as $m+n$, that form is used. The score is the confidence written in that trajectory. Black-box DTC still scores the same trajectory with the external auxiliaries under teacher forcing and does not use the generator's logits.

Verbalized confidence

Question: `[Question]`
Reason step-by-step to formulate your final answer. Your answer must be a mathematical value or short exact answer (e.g., an integer, fraction, radical, coordinate, or simplified expression), not a full sentence. If the question specifies a final report form (e.g., $m+n$ or $m-n$), use that form. Then, reason about the confidence in your answer. Conclude by providing a JSON object that states the final answer and your estimated confidence in it: `{`
`"final_answer": "Your final answer",`
`"confidence": "0-1"`
`}`

Figure 6: Prompt for Verbalized confidence. `[Question]` is replaced by the problem.

Verbalized TopK

Question: `[Question]`
Reason step-by-step to formulate 2 best guesses and probability that each is correct. Each answer must be a mathematical value or short exact answer (e.g., an integer, fraction, radical, coordinate, or simplified expression), not a full sentence. If the question specifies a final report form (e.g., $m+n$ or $m-n$), use that form. Your final output must be a JSON array: `[`

```
{
"candidate": "first most likely answer",
"confidence": "0-1"
},
{
"candidate": "second most likely answer",
"confidence": "0-1"
}
]
```

Figure 7: Prompt for Verbalized TopK with k=2. `[Question]` is replaced by the problem.

Verbalized probability distribution

Question: `[Question]`
Reason step-by-step to formulate your answer. You may propose multiple possible answers (fewer than five). Each answer must be a mathematical value or short exact answer (e.g., an integer, fraction, radical, coordinate, or simplified expression), not a full sentence. If the question specifies a final report form (e.g., $m+n$ or $m-n$), use that form. Always include "None of the above" as a possible answer. Reason about the confidence in each possible answer. Your final output must be a JSON array where the confidence scores form a probability distribution (they must sum to 1.0): `[`

```
{
"candidate": "Candidate 1",
"confidence": "0-1"
},
{
"candidate": "Candidate 2",
"confidence": "0-1"
},
...
{
"candidate": "None of the above",
"confidence": "0-1"
}
]
```

Figure 8: Prompt for Verbalized probability distribution. `[Question]` is replaced by the problem.

Verbalized confidence. We ask the model to write the answer and a confidence score in the same generation (Tian et al., 2023). The prompt is shown in Figure 6. A single stated number is easy to read, but the model can focus on the answer it already prefers.

Verbalized TopK. We ask the model to list k=2 candidate answers, each with a probability, and we take the probability of the highest-scoring candidate (Tian et al., 2023). The prompt is shown in Figure 7. Listing more than one candidate forces a comparison instead of a score for a single guess.

Verbalized probability distribution. We ask the model to assign probabilities to candidate answers that sum to 1. On open-ended questions the list includes "None of the above", which holds the remaining low-probability answers. The prompt is shown in Figure 8. The reported confidence is the probability of the selected answer, so that mass has to be shared with the other candidates.

B Complete Results

The tables below add the results omitted from the main text. Accuracy uses every sampled response. ECE and AUROC use the class-balanced sets in Appendix A.3. An arrow marks a row that only rescores the same trajectories, so that row has no accuracy of its own. A dash means the metric is missing or does not apply.

B.1 Full White-box Results

The main text reports white-box ECE (Table 1). Table 5 is the AUROC table for the same single-auxiliary setup, with the same generators, benchmarks, and baselines. PRM is a supervised reference and is not part of the unsupervised best and second-best comparison.

Table 5: Uncertainty quantification performance (AUROC, ↑ higher is better) in white-box settings. In each row, the best result is in **bold** and the second-best one is underlined (excluding PRM). †: UQAC variant by ours.

Reasoning Models	Acc	C_{NSL}	C_{mean}	Entropy Conf.	BaseCal	UQAC		Verb.	DTC (Ours)		PRM (ref.)
						attn	mean†		prod	lin	
MATH-500											
Qwen2.5-7B	76.1	67.0	66.5	61.4	58.5	59.2	62.0	74.7	80.4	**83.8**	95.4
Qwen2.5-14B	80.0	65.7	65.0	65.0	56.0	56.6	64.9	74.8	81.9	**85.0**	94.3
Qwen2.5-32B	82.4	66.7	65.8	65.5	50.6	59.7	65.0	83.8	82.7	**85.8**	94.4
Qwen3-8B	83.9	75.0	73.4	72.4	52.3	59.1	61.6	57.9	**79.3**	73.9	90.9
Qwen3-14B	86.5	74.3	72.8	72.0	53.2	66.1	75.6	61.8	**79.4**	75.6	89.8
Qwen3-32B	83.7	73.6	71.6	70.9	—	58.8	63.1	63.3	**80.6**	78.9	91.3
Gemma3-12B	84.9	61.9	60.3	62.7	40.9	61.9	58.1	**84.1**	81.9	84.0	93.6
Gemma3-27B	89.2	63.6	61.8	63.3	38.9	76.5	81.7	87.2	85.5	**87.9**	93.2
AMC23											
Qwen2.5-7B	53.6	72.4	71.8	67.9	66.6	63.9	68.1	67.1	**81.3**	81.2	90.8
Qwen2.5-14B	61.2	66.9	66.4	67.7	62.3	57.1	68.5	73.0	79.2	**80.3**	90.5
Qwen2.5-32B	66.4	67.1	66.3	69.8	60.8	62.2	68.6	73.7	77.6	**78.0**	87.2
Qwen3-8B	68.6	73.8	72.0	71.6	56.2	41.1	47.7	57.8	**78.7**	76.0	90.6
Qwen3-14B	74.1	73.5	71.5	72.6	56.6	66.1	72.9	68.8	**79.2**	75.3	89.5
Qwen3-32B	67.5	75.3	73.4	74.4	—	55.1	57.2	67.7	**80.3**	78.6	89.0
Gemma3-12B	66.8	69.7	68.6	70.0	54.2	61.2	61.9	78.7	**79.9**	78.7	86.6
Gemma3-27B	76.9	68.5	67.1	68.5	47.8	72.7	76.4	**85.6**	82.6	80.4	91.1
AIME24											
Qwen2.5-7B	12.6	79.0	78.0	74.4	75.2	55.1	61.5	79.0	**84.4**	75.6	96.2
Qwen2.5-14B	13.8	71.3	71.0	75.2	65.5	61.4	70.1	**77.8**	76.5	74.2	89.7
Qwen2.5-32B	16.9	75.6	75.2	79.3	72.2	67.5	73.5	**84.1**	80.0	70.4	89.6
Qwen3-8B	27.7	71.3	69.9	72.1	58.4	53.5	56.4	65.5	**74.1**	65.7	89.0
Qwen3-14B	27.1	70.8	69.4	70.6	61.4	64.9	72.6	63.9	**74.3**	67.0	88.6
Qwen3-32B	28.3	71.2	69.9	73.4	—	52.1	52.5	74.7	**76.2**	69.6	91.7
Gemma3-12B	23.9	81.0	80.1	79.8	71.1	64.8	60.9	74.8	**81.8**	66.4	93.8
Gemma3-27B	29.0	78.5	78.1	78.3	68.2	66.4	65.1	**87.8**	80.2	64.0	90.8
AIME25											
Qwen2.5-7B	9.1	**79.3**	78.9	77.8	78.7	67.7	71.1	69.6	76.5	64.3	88.6
Qwen2.5-14B	14.7	75.1	74.7	**75.9**	71.8	68.5	75.5	74.4	**75.9**	62.2	88.3
Qwen2.5-32B	12.2	79.0	79.1	**79.4**	75.4	52.4	77.5	72.1	76.7	60.1	83.8
Qwen3-8B	22.5	78.9	77.4	77.2	67.9	55.1	56.2	63.3	**82.4**	70.9	87.5
Qwen3-14B	27.5	82.5	81.3	82.8	73.8	67.7	84.1	52.1	**89.1**	81.8	87.5
Qwen3-32B	23.7	80.3	78.8	81.3	—	61.2	61.7	74.8	**86.8**	79.9	88.1
Gemma3-12B	18.8	84.9	84.1	84.4	69.0	64.8	66.7	85.7	**88.9**	76.3	89.6
Gemma3-27B	25.3	86.9	85.8	86.4	67.2	70.9	74.8	**91.3**	89.8	79.0	91.1
Average	48.0	73.8	72.7	73.2	61.8	61.6	66.7	73.5	**80.8**	75.3	90.4

B.2 Full Black-box Results

Tables 6 and 7 are the Qwen3-4B and Qwen3-30B-A3B results. Each table gives accuracy for the original generation protocol, plus ECE and AUROC on AIME24, AIME25, HMMT25, and HMMT26, with averages. DTC and PRM rescore the frozen Default CoT trajectories. Within a confidence protocol, that protocol and its DTC rows use the same trajectories.

Table 6: Uncertainty quantification performance with Qwen3-4B in black-box settings.

Method	AIME24			AIME25			HMMT25			HMMT26		
	Acc↑	ECE↓	AUC↑	Acc↑	ECE↓	AUC↑	Acc↑	ECE↓	AUC↑	Acc↑	ECE↓	AUC↑
Default CoT	61.4	–	–	45.6	–	–	30.3	–	–	34.7	–	–
→ + PRM	–	36.0	**82.9**	–	36.6	80.3	–	37.8	53.3	–	33.4	78.0
→ + DTC_{lin}	–	**11.3**	72.9	–	**9.0**	81.5	–	**13.1**	69.6	–	**13.0**	74.6
→ + DTC_{prod}	–	16.3	75.8	–	10.7	**84.8**	–	15.4	**78.0**	–	14.3	**79.0**
Verb. Conf.	61.8	45.6	**80.8**	46.6	45.9	**85.6**	27.3	45.6	**81.9**	33.0	45.1	79.8
→ + DTC_{lin}	–	**11.9**	71.6	–	**9.9**	80.2	–	16.5	66.5	–	12.9	77.1
→ + DTC_{prod}	–	17.4	74.7	–	14.3	84.4	–	**16.1**	73.7	–	**11.6**	**81.3**
Verb. TopK	52.4	47.0	**68.3**	42.0	47.0	75.3	26.4	46.8	**74.0**	33.3	45.5	**77.3**
→ + DTC_{lin}	–	**19.3**	60.6	–	**16.2**	74.0	–	**22.1**	54.2	–	**9.9**	70.9
→ + DTC_{prod}	–	30.1	61.8	–	31.4	**79.0**	–	30.2	62.7	–	25.0	75.1
Verb. PD	59.3	46.9	69.9	46.9	48.1	73.7	26.7	43.9	**75.5**	32.5	42.9	70.9
→ + DTC_{lin}	–	**14.4**	68.7	–	**9.4**	76.3	–	**14.8**	65.6	–	**7.2**	77.0
→ + DTC_{prod}	–	25.7	**72.6**	–	22.4	**82.1**	–	22.7	**75.5**	–	20.1	**82.2**

Table 7: Uncertainty quantification performance with Qwen3-30B-A3B in black-box settings.

Method	AIME24			AIME25			HMMT25			HMMT26		
	Acc↑	ECE↓	AUC↑	Acc↑	ECE↓	AUC↑	Acc↑	ECE↓	AUC↑	Acc↑	ECE↓	AUC↑
Default CoT	73.4	–	–	59.8	–	–	42.4	–	–	43.5	–	–
→ + PRM	–	37.5	**86.0**	–	37.8	78.4	–	38.7	53.6	–	34.6	77.4
→ + DTC_{lin}	–	**15.9**	70.2	–	**12.3**	75.8	–	**13.1**	72.1	–	**9.6**	77.8
→ + DTC_{prod}	–	18.2	75.1	–	16.6	**79.6**	–	19.5	**78.7**	–	17.1	**80.5**
Verb. Conf.	75.7	44.5	**83.9**	60.6	44.3	**80.8**	41.8	42.9	**85.3**	43.9	43.9	79.2
→ + DTC_{lin}	–	**14.0**	73.7	–	**13.2**	73.7	–	**11.4**	74.3	–	**10.7**	78.5
→ + DTC_{prod}	–	20.6	78.4	–	18.4	77.8	–	19.3	80.5	–	16.6	**81.4**
Verb. TopK	71.1	40.2	**88.2**	55.8	40.6	**88.7**	40.7	41.2	**85.4**	43.4	41.6	**83.0**
→ + DTC_{lin}	–	**36.4**	44.3	–	**35.2**	49.0	–	**22.4**	64.5	–	**21.9**	66.6
→ + DTC_{prod}	–	46.7	45.2	–	47.0	50.2	–	32.9	71.0	–	37.0	69.0
Verb. PD	75.3	33.5	**75.5**	62.6	34.1	69.6	41.0	31.1	**77.3**	43.7	34.9	72.3
→ + DTC_{lin}	–	**18.8**	63.0	–	**16.8**	72.0	–	**17.9**	68.4	–	**15.3**	77.3
→ + DTC_{prod}	–	31.7	67.3	–	29.4	**75.5**	–	29.2	76.3	–	29.3	**80.6**

B.3 Verbalized Confidence with DTC

Tables 8 and 9 repeat the verbalized-confidence experiment on Qwen3-30B-A3B and Qwen3-4B. For each protocol, the DTC adjustment uses the black-box DTC_{prod} setup and the same frozen trajectory as the unadjusted verbalized score, so accuracy does not change.

Table 8: Verbalized vs. w/ DTC on Qwen3-30B-A3B.

Method	AIME24		AIME25		HMMT25		HMMT26	
	ECE↓	AUC↑	ECE↓	AUC↑	ECE↓	AUC↑	ECE↓	AUC↑
Verb. Conf.	44.5	83.9	44.3	80.8	42.9	85.3	43.9	79.2
w/ DTC	17.8	87.0	15.7	83.8	13.0	87.9	17.4	85.9
Verb. TopK	40.2	88.2	40.6	88.7	41.2	85.4	41.6	83.0
w/ DTC	11.1	83.0	10.4	84.9	5.4	86.5	6.1	86.4
Verb. PD	33.5	75.5	34.1	69.6	31.1	77.3	34.9	72.3
w/ DTC	23.8	79.7	23.4	76.9	22.1	80.0	20.1	79.5

Table 9: Verbalized vs. w/ DTC on Qwen3-4B.

Method	AIME24		AIME25		HMMT25		HMMT26	
	ECE↓	AUC↑	ECE↓	AUC↑	ECE↓	AUC↑	ECE↓	AUC↑
Verb. Conf.	45.6	80.8	45.9	85.6	45.6	81.9	45.1	79.8
w/ DTC	30.8	82.0	29.4	87.0	31.0	81.6	31.1	81.2
Verb. TopK	47.0	68.3	47.0	75.3	46.8	74.0	45.5	77.3
w/ DTC	33.7	68.6	32.8	75.5	34.3	73.8	30.9	76.9
Verb. PD	46.9	69.9	48.1	73.7	43.9	75.5	42.9	70.9
w/ DTC	33.6	72.8	35.7	76.6	30.3	77.2	29.9	73.4

C Additional Analyses

C.1 Black-box Auxiliary Size

For dual-auxiliary $\mathrm{DTC}_{\mathrm{lin}}$, DeepSeek-V3.2 generates the trajectories and both auxiliaries are Qwen2.5 models (Figure 9). On AIME25 we search θ in steps of 0.05 and report the θ with the lowest ECE. The lowest ECE is similar across auxiliary pairs.

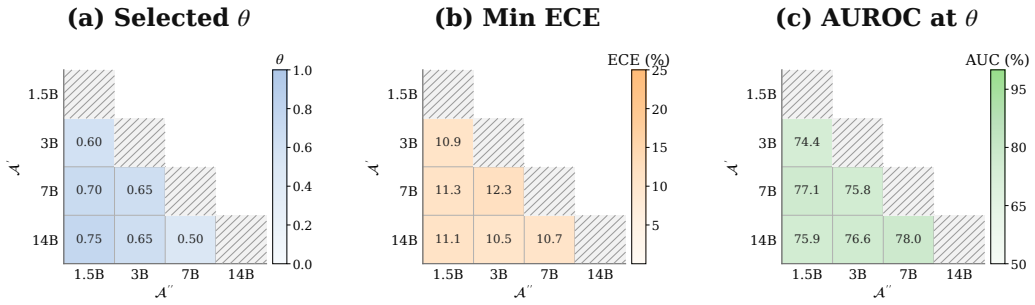

Figure 9: Black-box sensitivity of $\mathrm{DTC}_{\mathrm{lin}}$ to θ and auxiliary size. We search θ in steps of 0.05 and report the θ (a) with the lowest ECE (b) and its AUROC (c). Cells without a finished probe are left blank.

C.2 Black-box Threshold Curves

With $\mathcal{A}''$ fixed at 1.5B, we vary θ for black-box $\mathrm{DTC}_{\mathrm{lin}}$ on DeepSeek-V3.2 trajectories (Figure 10). The main experiments use θ=0.70, where ECE is still low.

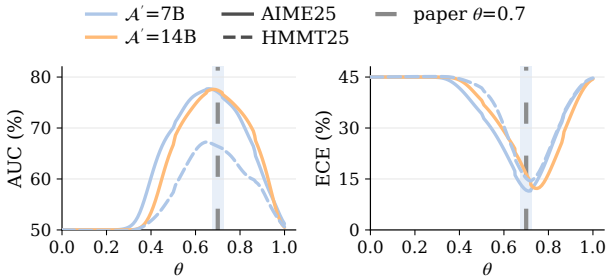

Figure 10: $\mathrm{DTC}_{\mathrm{lin}}$ sensitivity to θ with $\mathcal{A}''$ fixed at 1.5B (DeepSeek-V3.2 generating; solid: AIME25; dashed: HMMT25). The shaded band marks the black-box operating point θ=0.70.

C.3 Count and Ratio

Longer reasoning traces are often less accurate, and recent work links that drop to overthinking (Ghosal et al., 2025; Gema et al., 2025). We check whether the divergent-token count is only a

proxy for this length effect. Let L be the number of reasoning tokens on a path. The count grows with the path, so the two panels of Figure 11 plot

$$m = |T_{\mathrm{div}}(\theta)|, \qquad \frac{m}{L} = \frac{|T_{\mathrm{div}}(\theta)|}{L}. \tag{16}$$

The runs use the single-auxiliary setting of Figure 1(c): MATH-500, Qwen2.5-7B, auxiliaries 1.5B, 3B, 14B, and 32B, each at the threshold used there. Accuracy uses every sampled answer, including truncated generations.

On the left of Figure 11, accuracy falls from about 0.97–0.99 at a count of zero to about 0.24–0.29 once the count reaches 11 or more. On the right, we group the same trajectories into ten quantile bins of the ratio. Accuracy still falls, from about 0.97–0.99 in the lowest bin to about 0.45–0.54 in the highest. The correlation of these bin means is $r = -0.985$ for the count and $r = -0.937$ for the ratio. Dividing by path length does not remove the decline, so the count is not only a proxy for how long the path is.

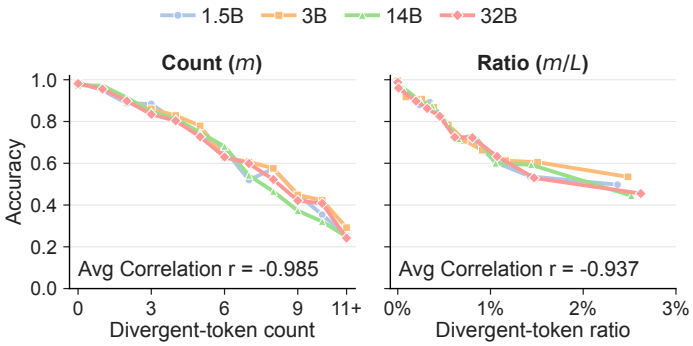

Figure 11: Divergent-token count and ratio against path accuracy, single-auxiliary Qwen2.5-7B on MATH-500. Each curve is one auxiliary at the threshold used in Figure 1(c).

C.4 Choice of Divergence Measure

We keep the count estimator and the auxiliaries fixed, and compare JSD with forward KL and reverse KL (Figures 12 and 13). White-box runs use a 1.5B auxiliary and generators of 7B, 14B, and 32B. Black-box runs use Qwen2.5-7B and Qwen2.5-1.5B. ECE against θ is U-shaped for all three scores, and AUROC rises where ECE falls. KL needs a larger θ than JSD, and that θ is less stable from model to model. JSD is steadier near the θ used in the main experiments.

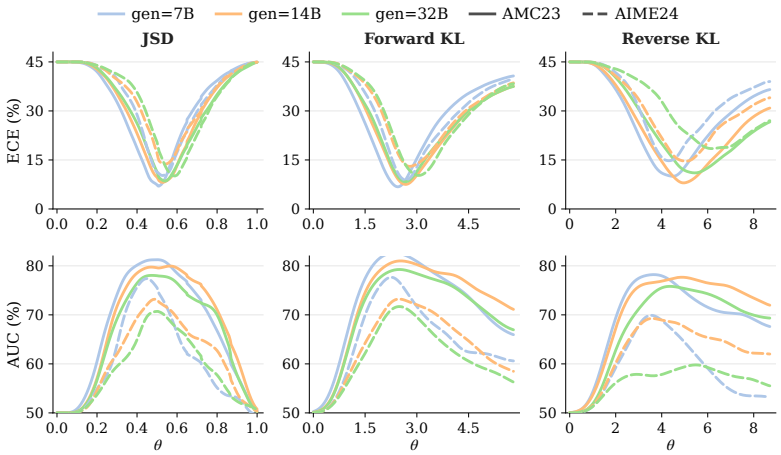

Figure 12: Divergence choice for divergent-token selection. Each column is one disagreement score (JSD, forward KL, reverse KL) for $\mathrm{DTC}_{\mathrm{lin}}$ with auxiliary fixed at 1.5B and generating models $\in \{7, 14, 32\}$B (solid: AMC23; dashed: AIME24). The top row shows ECE (%) and the bottom row AUROC (%) versus θ.

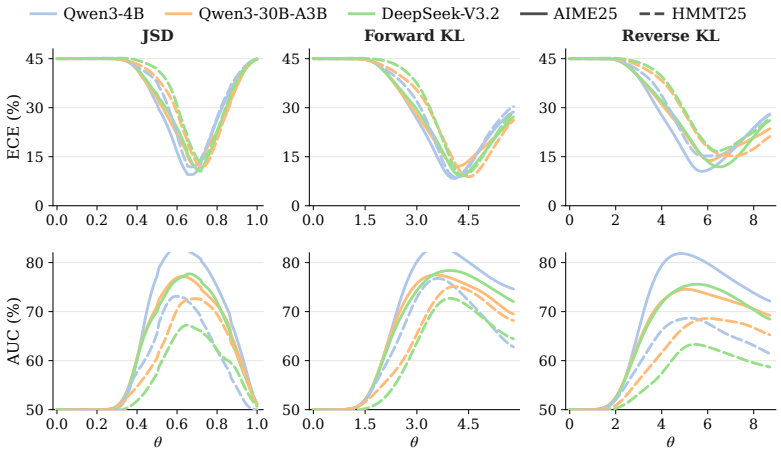

Figure 13: Divergence choice for black-box divergent-token selection. Each column is one disagreement score (JSD, forward KL, reverse KL) for $\mathrm{DTC}_{\mathrm{lin}}$ with auxiliaries fixed at Qwen2.5-7B and Qwen2.5-1.5B and generators Qwen3-4B, Qwen3-30B-A3B, and DeepSeek-V3.2 (solid: AIME25; dashed: HMMT25). The top row shows ECE (%) and the bottom row AUROC (%) versus θ.

C.5 Threshold and Saturation Sensitivity

In the white-box setting, Figure 14 varies θ and the saturation length n of $\mathrm{DTC}_{\mathrm{lin}}$. Each panel uses the smallest generator in its family: Qwen2.5-7B, Qwen3-8B, or Gemma3-12B. The curves average MATH-500, AMC23, AIME24, and AIME25. The ECE minimum moves with the family, but the θ values from the main text are still where ECE is low, and the comparison across n supports $n = 10$.

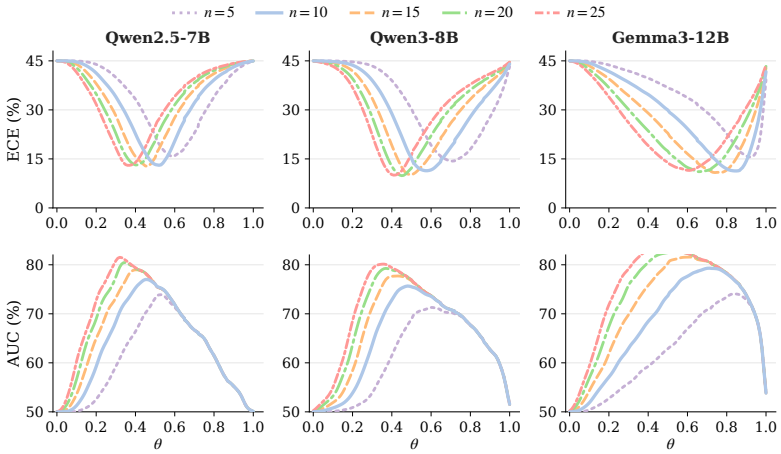

Figure 14: Saturation length n for $\mathrm{DTC}_{\mathrm{lin}}$. The top row shows ECE (%) and the bottom row AUROC (%) versus θ at step 0.01, averaged over MATH-500, AMC23, AIME24, and AIME25.

C.6 Black-box Saturation Length

Black-box $\mathrm{DTC}_{\mathrm{lin}}$ uses DeepSeek-V3.2, Qwen3-4B, and Qwen3-30B-A3B, with Qwen2.5 auxiliaries of 7B and 1.5B (Figure 15). Averaged over AIME24, AIME25, and HMMT25, these curves also support $n = 10$.

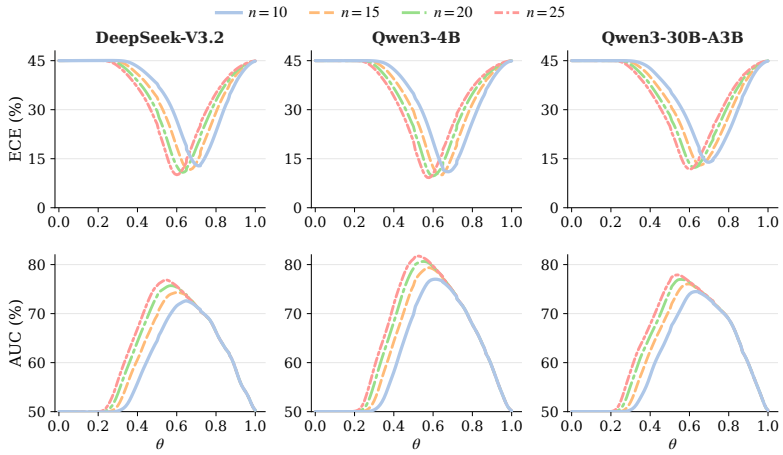

Figure 15: Saturation length n for black-box $\mathrm{DTC}_{\mathrm{lin}}$. Each column is one generator with Qwen2.5 auxiliaries 7B and 1.5B. The top row shows ECE (%) and the bottom row AUROC (%) versus θ at step 0.01, averaged over AIME24, AIME25, and HMMT25.

C.7 Offset k for $\mathrm{DTC}_{\mathrm{prod}}$

Figure 16 varies k in $\mathrm{DTC}_{\mathrm{prod}}$ on the smallest white-box generator in each family. ECE is averaged over MATH-500, AMC23, AIME24, and AIME25. ECE against θ stays U-shaped, and the minimum moves with k and with the family. At the θ used in the main text, $k = 4$ still gives low ECE, so we keep $k = 4$.

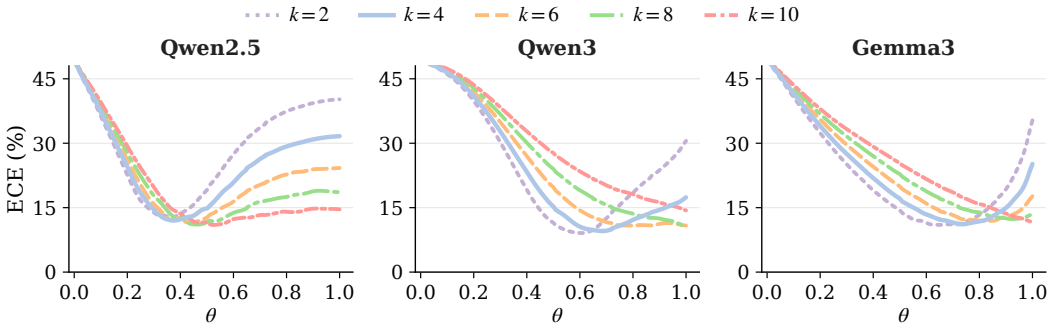

Figure 16: Offset k for $\mathrm{DTC}_{\mathrm{prod}}$. Each column is the smallest white-box generator in its family (Qwen2.5-7B, Qwen3-8B, Gemma3-12B) with the family auxiliary. ECE (%) is shown versus θ at step 0.01, averaged over MATH-500, AMC23, AIME24, and AIME25.