Inference Time Optimization with Confidence Dynamics

Yu Wang, Minghao Liu, Jiayun Wang, Jinrui Huang, Shah Ankit Parag, Wei Wei

Accenture Advanced AI Research Center

Contacts:
Yu Wang
feather1014@gmail.com

What is inference time optimization/scaling?

At inference time, determine the optimal answer from multiple sampled paths/trajectory/traces (Brown, et al., 2024, Snell et al., 2024, figure from Wang et al., 2023)

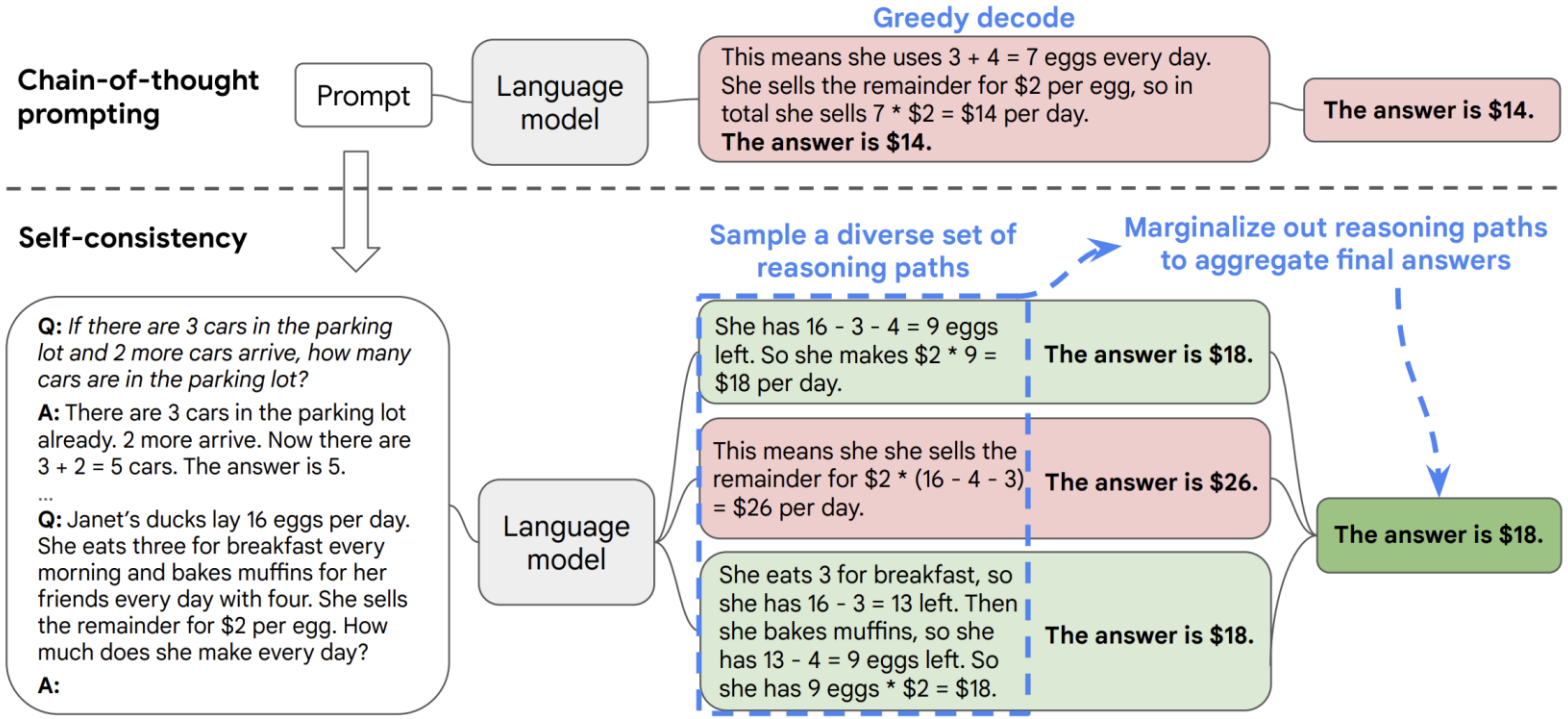


Figure 1: The self-consistency method contains three steps: (1) prompt a language model using chain-of-thought (CoT) prompting; (2) replace the "greedy decode" in CoT prompting by sampling from the language model's decoder to generate a diverse set of reasoning paths; and (3) marginalize out the reasoning paths and aggregate by choosing the most consistent answer in the final answer set.

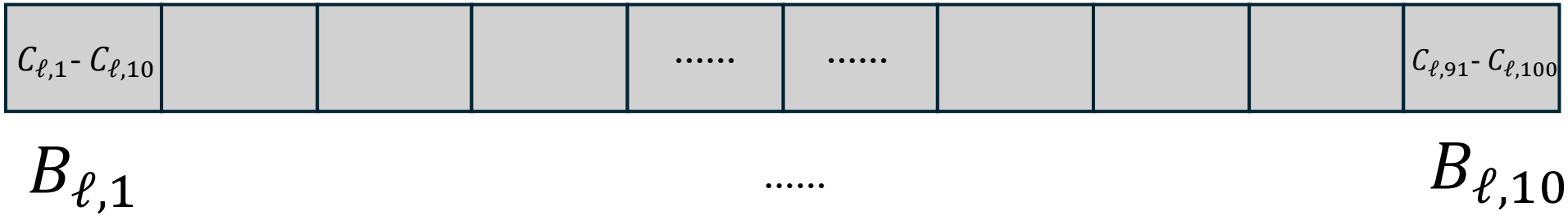
Snell., et al, Scaling LLM test-Time compute optimally can be more effective than scaling model parameters
Brown, et al., Large Language Monkeys: Scaling Inference Compute with Repeated Sampling, 2024
Wang et al, Self-Consistency Improves Chain of Thought Reasoning in Language Models, 2023

Question:

Confidence dynamic along the trajectory?

We partition the answer sequence into N equal-sized bins， where the bin size is b= T/N

Position-Normalized Binning: N=10 Bins (T=100 tokens example)



Question:

Confidence dynamic (some stats) along the trajectory?

Observation:

Correct answer traces tend to exhibit confidence improvement over time (positive confidence gain)

Incorrect traces show attenuated or declining confidence as reasoning proceeds.

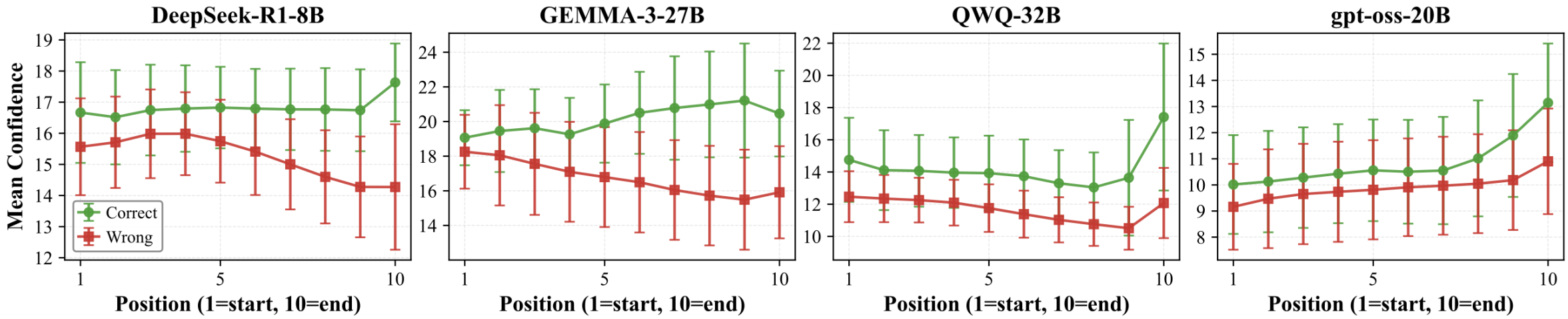


Figure 2. Confidence dynamic curves for the reasoning trajectories (partitioned in 10 bins) for different models on AIME 2025 dataset. Significance tests (Table 4 in Appendix) show that correct traces have greater confidence gains than wrong traces.

We compute the Confidence Dynamic Gain (CDG) as: the difference between tail bin and head bin confidence

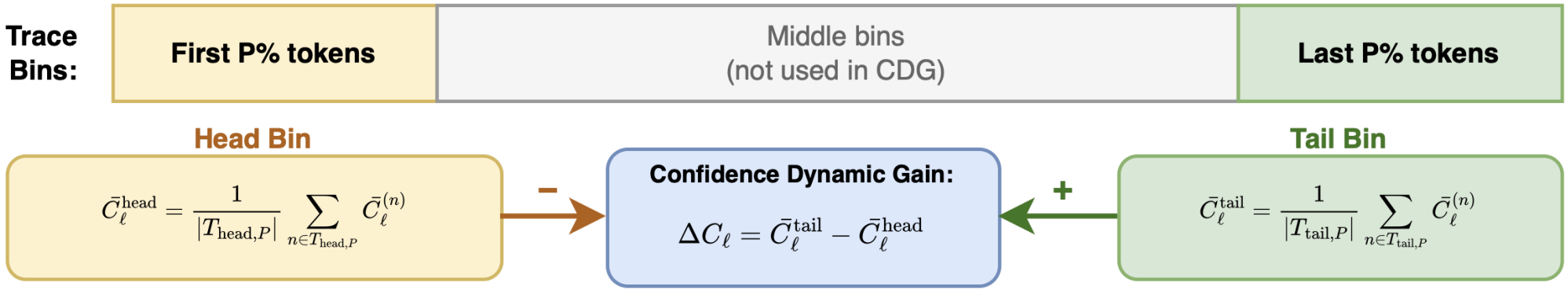


Figure 1. Confidence Dynamic Gain (CDG) computation, i.e., ΔC_ℓ. Each trace with T tokens is partitioned into N position-normalized bins. The head bin confidence $\bar{C}_\ell^{\text{head}}$ averages over the first $P\%$ of positions, and the tail bin confidence $\bar{C}_\ell^{\text{tail}}$ averages over the last $P\%$. The CDG ΔC_ℓ captures how confidence evolves from reasoning start to conclusion.

$$\Delta C_\ell = \frac{1}{|T_{\text{tail},P}|} \sum_{n \in T_{\text{tail},P}} \bar{C}_\ell^{(n)} - \frac{1}{|T_{\text{head},P}|} \sum_{n \in T_{\text{head},P}} \bar{C}_\ell^{(n)},$$

$\Delta C_\ell > 0$ **gaining confidence as the reasoning chain progresses to final answer**

$\Delta C_\ell < 0$ **losing confidence as the reasoning chain progresses to final answer**

Empirical Result: Benchmark

Table 1. Accuracy (%) of different voting methods across models and benchmarks. Best average results are in **bold**. Pass@1 is the single-trace baseline. For CDG (ours), we use $\alpha = 0.5$ and $P\% = 10\%$ across all experiments. $\beta = 10$ for DeepSeek-R1-8B and gpt-oss-20B; $\beta = 3$ for Gemma-3-27B and QWQ-32B. "Majority" is for Majority Vote (Wang et al., 2022). D-CDG is for "Degenerated-CDG", i.e., an ablation of CDG with either $\beta = 0, \alpha = 0.5$ (no confidence gain dynamics, but with count dampening) or $\alpha = 1, \beta \neq 0$ (no count dampening, but with confidence dynamic gain). DC-Mean/DC-Tail are DeepConf-Mean/Tail with top 10% filtering (Fu et al., 2025).

Model	**Dataset**	**Pass@1**	**Majority**	**DC-Mean**	**DC-Tail**	**D-CDG** ($\alpha = 1$)	**D-CDG** ($\beta = 0$)	**CDG (ours)**
DeepSeek-R1-8B	AIME 2024	86.5	90.0	90.0	93.3	93.3	90.0	93.3
	AIME 2025	77.5	83.3	83.3	83.3	90.0	83.3	93.3
	BRUMO 2025	79.7	93.3	93.3	93.3	93.3	93.3	93.3
	HMMT 2025	59.5	70.0	70.0	83.3	76.7	70.0	83.3
	Average	75.8	84.2	84.2	88.3	88.3	84.2	**90.8**
Gemma-3-27B	AIME 2024	31.5	50.0	50.0	53.3	56.7	50.0	56.7
	AIME 2025	24.7	30.0	30.0	46.7	33.3	30.0	40.0
	BRUMO 2025	35.9	40.0	40.0	46.7	46.7	43.3	46.7
	HMMT 2025	10.8	20.0	20.0	13.3	23.3	20.0	23.3
	Average	25.7	35.0	35.0	40.0	40.0	35.8	**41.7**
gpt-oss-20B	AIME 2024	77.9	93.3	93.3	93.3	93.3	93.3	93.3
	AIME 2025	71.1	90.0	90.0	90.0	90.0	90.0	93.3
	BRUMO 2025	64.8	80.0	83.3	83.3	83.3	83.3	83.3
	HMMT 2025	52.2	66.7	70.0	73.3	70.0	70.0	73.3
	Average	66.5	82.5	84.2	85.0	84.2	84.2	**85.8**
QwQ-32B	AIME 2024	81.7	86.7	90.0	86.7	90.0	90.0	90.0
	AIME 2025	71.6	76.7	76.7	76.7	76.7	76.7	76.7
	BRUMO 2025	76.8	80.0	80.0	86.7	86.7	80.0	90.0
	HMMT 2025	48.7	56.7	56.7	63.3	63.3	56.7	63.3
	Average	69.7	75.0	75.9	78.3	79.2	75.8	**80.0**
	Overall Average	59.4	69.2	69.8	72.9	72.9	70.0	**74.6**

Empirical Result: Ablation Study

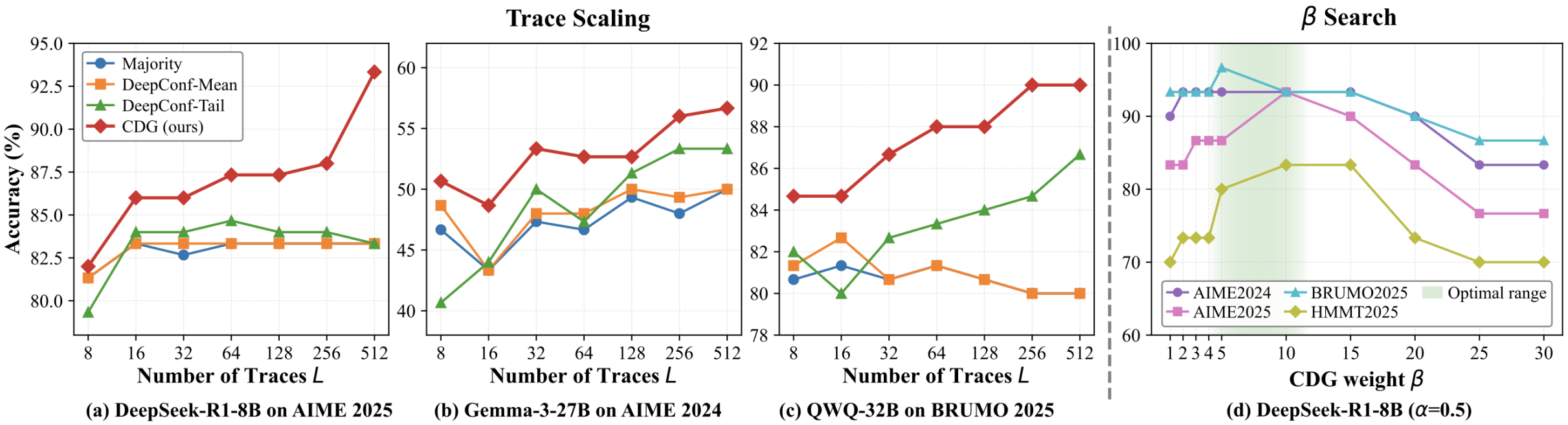


Figure 3. Ablation study of trace number (budget) L and CDG weight β. (a) Accuracy vs L for DeepSeek-R1-8B on AIME 2025; (b) Accuracy vs L for Gemma-3-27B on AIME 2024. (c) Accuracy vs L for QwQ-32B on BRUMO 2025. (d)Accuracy vs β for CDG using DeepSeek-R1-8B across 4 datasets. The green band highlights the estimated optimal range of $\beta \in [0.5r_b, 1.5r_b]$ as discussed in Section 3.3. See Figures 5 and 6 in the Appendix for complete results.

Empirical Result: Statistics

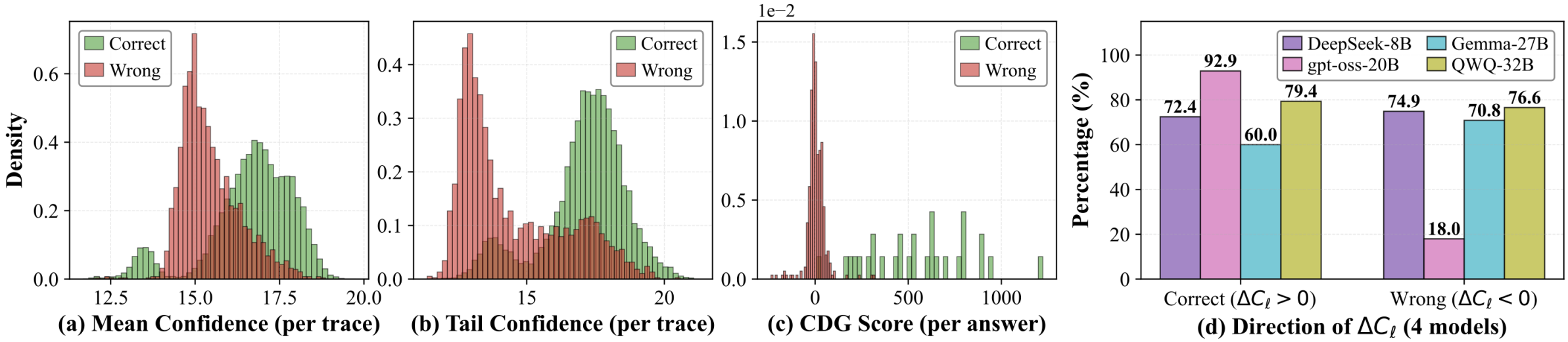


Figure 4. Confidence metric statistical analysis on AIME 2024 using DeepSeek-R1-8B. Wrong traces are marked in red while correct ones in green.(a) Mean token confidence distribution. (b) Tail token confidence (last 10%). (c) CDG score (d) Direction analysis for ΔC_ℓ.

Theoretical Analysis

Theorem 4.6 (Confidence Gain Separation). *Under GRPO training with verifiable rewards and Assumptions 4.3-4.4, the expected confidence dynamic gain satisfies:*

$$\mathbb{E}[\Delta C_\ell \mid \textit{Correct}] - \mathbb{E}[\Delta C_\ell \mid \textit{Incorrect}] \tag{12}$$

$$\geq c \cdot \eta_{\textit{eff}} \cdot \sqrt{k(G-k)} \cdot \left(\gamma M - \frac{1}{M}\right) > 0, \tag{13}$$

for constants $c > 0$ and effective learning rate $\eta_{\textit{eff}}$. The separation is positive provided $\gamma > 1/M^2$.

For correct traces,
diverse reasoning paths tend to collapse into the same answer, creating a high tail-to-head concentration ratio M. This amplifies the positive reinforcement at the tail.

For incorrect traces,
because during training answer errors are inherently more diverse than the unique ground truth, received less reinforcement at the tail.

Inference Time Optimization with Confidence Dynamics

Yu Wang, Minghao Liu, Jiayun Wang, Jinrui Huang, Shah Ankit Parag, Wei Wei

Accenture Advanced AI Research Center

Contacts:
Yu Wang
feather1014@gmail.com