

CAT: Confidence-Adaptive Thinking for Efficient Reasoning of Large Reasoning Models

Qizhi Jiang¹ Shuo Wang¹ Pei Ke^{1,2,*} Yuhang Song¹ Ke Qin^{1,2}

¹Laboratory of Intelligent Collaborative Computing,

University of Electronic Science and Technology of China, Chengdu, China

²Ubiquitous Intelligence and Trusted Services Key Laboratory of Sichuan Province

{jiangqizhi, 202422900227}@std.uestc.edu.cn, kepei@uestc.edu.cn

songyuhang@std.uestc.edu.cn, qinke@uestc.edu.cn

Abstract

Large Reasoning Models (LRMs) have achieved remarkable success on complex tasks by leveraging long chain-of-thought (CoT) trajectories, yet they frequently exhibit overthinking on simple queries, resulting in significant token overhead and reduced inference efficiency. However, existing compression methods predominantly apply uniform length reduction or rely on coarse-grained difficulty estimation, often leading to performance degradation on difficult problems. To address this limitation, we propose Confidence-Adaptive Thinking (CAT), a framework that incorporates the model’s intrinsic self-certainty signals as confidence into the preference optimization process, which autonomously modulates reasoning lengths based on problem difficulty. Experimental results show that CAT consistently outperforms state-of-the-art baselines on reasoning accuracy across multiple benchmarks on different base models. Our work enables LRMs to effectively compress confident responses while deliberating on uncertain ones, offering a potentially robust solution for balancing accuracy and latency in practical industrial scenarios.

1 Introduction

Recently, large reasoning models (LRMs) have rapidly emerged and made substantial progress on complex natural language processing (NLP) tasks, as exemplified by OpenAI-o1 (OpenAI, 2024) and DeepSeek-R1 (DeepSeek-AI, 2025). These models are equipped with the ability to generate long reasoning chains, demonstrating strong potential on challenging reasoning problems such as mathematical competitions (Xu et al., 2025). However, while LRMs heavily rely on long chain-of-thought (CoT) traces to perform well on difficult tasks, they tend

to produce redundant reasoning and self-reflection for simple inputs, incurring pronounced overthinking and token overhead (Chen et al., 2024; Feng et al., 2025; Liu et al., 2025; Sui et al., 2025). This behavior leads to verbose thought chains that increase computation cost and reduce overall inference efficiency. Accordingly, how to enable LRMs to dynamically adjust token consumption based on the input difficulty has attracted increasing attention, determining the practical industrial usability of LRMs in terms of the balance between accuracy and latency (Shen et al., 2025a).

Most of the existing approaches focus on reasoning compression and length control predominantly, which treat shortening reasoning chains as the primary objective (Qu et al., 2025) and apply a uniform reduction of reasoning tokens to all the queries (Xia et al., 2025; Chen et al., 2024; Ma et al., 2025; Munkhbat et al., 2025). While such methods can substantially decrease generation length, they often incur non-trivial performance degradation on difficult problems, since complex tasks still require sufficient reasoning depths and lengths to sustain accurate answers (Muennighoff et al., 2025; Zeng et al., 2024). Another line of work resorts to difficulty-adaptive reasoning to mitigate the imbalance between overthinking for easier instances and underthinking for harder ones. This category of methods tends to dynamically adjust the budget of output tokens based on the model performance (Shen et al., 2025a).

However, existing works on adaptive reasoning still face a severe challenge of coarse-grained difficulty estimation. Current methods utilize the accuracy of model outputs to measure the problem difficulty and roughly determine the output length (Shen et al., 2025a). We argue that this coarse-grained estimation heavily relies on external labels and provides a partial assessment merely on the answer, rather than measuring the quality of the whole reasoning chains generated by LRMs.

*Corresponding author.

To address this limitation, we propose CAT (Confidence-Adaptive Thinking), an adaptive reasoning framework driven by the model’s intrinsic confidence. Inspired by recent works on the quality estimation from the model’s internal token distributions (Fu et al., 2025; Geng et al., 2024; Fadeeva et al., 2024), our main idea is to leverage self-certainty (Kang et al., 2025) as the intrinsic fine-grained indicator to distinguish high-quality reasoning trajectories from erroneous ones. **Firstly**, CAT employs self-certainty as the model’s intrinsic confidence metric to estimate the quality of generated reasoning trajectories, which reflects the problem difficulty. Based on the separation of confidence and lengths between different trajectories, we further construct preference data to make the model capture the relationship between problem difficulties and output lengths. **Secondly**, we devise a confidence-weighted preference optimization (CWPO) method, which weights the vanilla preference optimization objective with confidence. This encourages the model to compress reasoning steps under high confidence while retaining necessary exploration otherwise, thereby mitigating overthinking for simple cases and maintaining reasoning performance especially for hard ones.

In summary, our main contributions are¹:

- We introduce the confidence-adaptive thinking (CAT) framework that shifts the paradigm of efficient reasoning from external supervision to intrinsic confidence awareness. CAT enables reasoning models to autonomously perceive problem difficulty and modulate their thinking depth.
- We propose the confidence-weighted preference optimization (CWPO) objective that dynamically weights the vanilla objective based on the calibration ratio of confidence to length. CWPO mitigates overthinking while preserving the model’s ability to explore complex reasoning paths if necessary.
- We conduct extensive experiments across three challenging benchmarks and show superior performance of CAT over state-of-the-art baselines on the balance between inference efficiency and reasoning accuracy.

¹Our codes are available at <https://github.com/Jiang9732/CAT-code>.

2 Related Work

Efficient Reasoning in LRMs. Recent studies have increasingly focused on the phenomenon of overthinking in large reasoning models (Sui et al., 2025; Wu et al., 2025; Wang et al., 2025). Existing efficient reasoning methods can generally be categorized into two streams. The first involves training strategies to equip LRMs with the ability to generate concise reasoning chains, spanning from supervised fine-tuning (Cui et al., 2025; Xia et al., 2025) to reinforcement learning (Shen et al., 2025a; Aggarwal and Welleck, 2025; Luo et al., 2025; Yu et al., 2025). The second category comprises inference-time methods, including prompting (Han et al., 2025; Renze and Guven, 2024; Nayab et al., 2024), task routing (Chuang et al., 2025; Ong et al., 2025), latent space compression (Hao et al., 2024; Shen et al., 2025b), and dynamic decoding (Sun et al., 2024; Zhang, 2025).

Compared with existing works on training methods of efficient reasoning, our work utilizes the model’s confidence as the estimation of problem difficulty, instead of solely depending on external reward models and extrinsic metrics. This makes the full usage of the model’s intrinsic property to achieve adaptive reasoning.

Confidence Utilization in LRMs. Recent works have shown that the model confidence potentially indicate the quality of reasoning chains (Fu et al., 2025; Geng et al., 2024; Kang et al., 2025; Fadeeva et al., 2024). As one of the representative metrics to reflect confidence, self-certainty (Kang et al., 2025) has been primarily applied to Best-of-N selection (Fu et al., 2025). For comparison, our work uses self-certainty as the model’s confidence to self-evaluate the quality of generated reasoning chains, which guides the adaptive thinking via preference optimization, instead of merely injecting it into the inference stage.

3 Methodology

3.1 Task Definition and Method Overview

Given an input question x , our goal is to acquire a reasoning trajectory y that contains a multi-step reasoning process and a final answer. Under the precondition of accuracy, y is required to become short for simple problems while being long for hard ones if necessary.

An overview of our framework is presented in Figure 1. Firstly, we sample multiple reasoning trajectories for each question and compute their

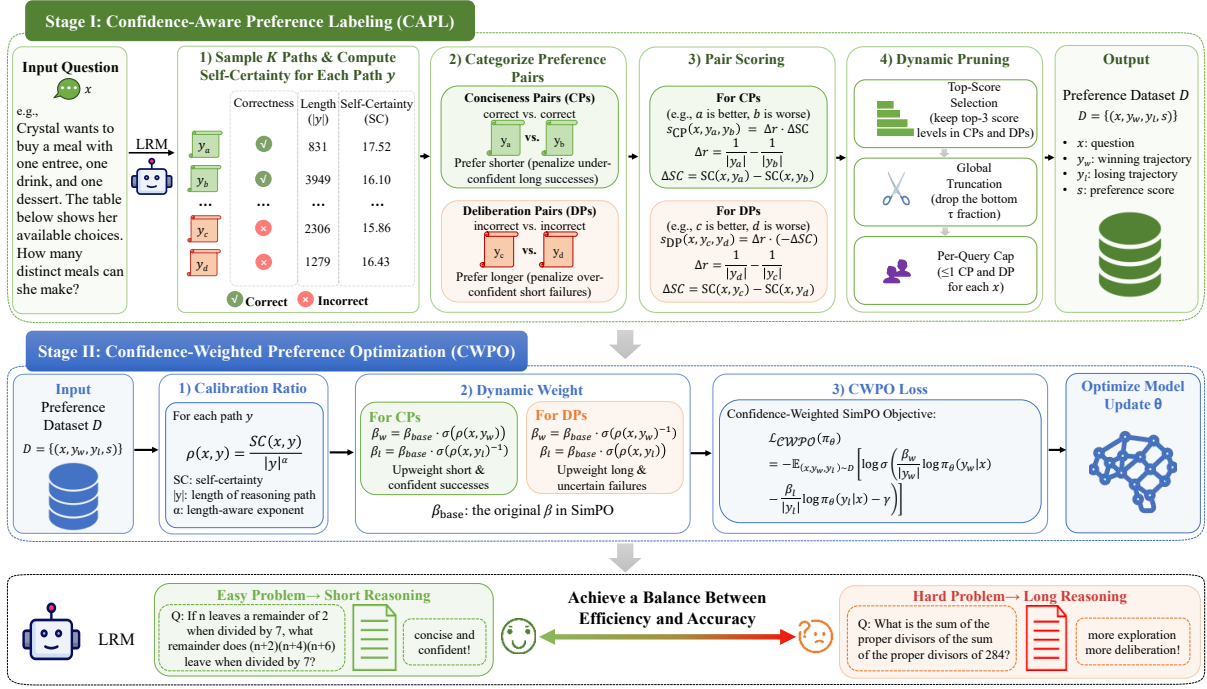


Figure 1: Overview of the CAT framework.

path-level self-certainty scores as confidence via a dedicated forward pass (Section 3.2.1). Secondly, we construct preference pairs based on the confidence and lengths, and apply dynamic selection to prioritize more informative supervision (Section 3.2.1). Finally, we fine-tune the base LRM with a confidence-weighted preference optimization objective, which incorporates confidence and lengths to further modulate the preference strength, achieving conditional length regulation (Section 3.2.2).

3.2 Confidence-Adaptive Thinking

Our confidence-adaptive thinking framework consists of two stages, including confidence-aware preference labeling and confidence-weighted preference optimization. While the first stage aims to incorporate confidence as intrinsic signals to construct fine-grained preference pairs, the second stage further utilizes confidence to further enhance the preference optimization objective.

3.2.1 Confidence-Aware Preference Labeling

To build the preference dataset, we first sample K reasoning trajectories $\{y^{(k)}\}_{k=1}^K$ for the question x from the base reasoning model, each of which is a token sequence $y^{(k)} = (y_1^{(k)}, \dots, y_{n_k}^{(k)})$ with the length of n_k . The goal of this stage is to construct a preference dataset $\mathcal{D} = \{(x, y_w, y_l, s)\}$, where y_w and y_l denote the winning and losing trajectories

for the same input x , and s indicates the confidence-calibrated preference score.

Self-Certainty as Intrinsic Confidence. To capture the model’s intrinsic confidence during reasoning, we follow Kang et al. (2025) to employ self-certainty, which can also serve as a trajectory-level quality measure. Formally, assuming that $p_\theta(\cdot | x, y_{\leq i})$ denotes the next-token distribution at i -th position, V indicates the vocabulary size, and \mathcal{U} represents the uniform distribution over V , self-certainty (SC) can be computed as follows:

$$SC(x, y) = -\frac{1}{nV} \sum_{i=1}^n \sum_{j=1}^V \log \left(V \cdot p_\theta(j | x, y_{\leq i}) \right) \quad (1)$$

which corresponds to measuring the KL divergence $D_{KL}(\mathcal{U} \| p_\theta(\cdot | x, y_{\leq i}))$ and averaging this quantity over i . Intuitively, a larger divergence from the uniform distribution implies a more peaked (and thus more certain) predictive distribution, leading to higher SC. Conversely, a distribution closer to uniform is flatter, indicating greater uncertainty and yielding lower SC.

Preference Pair Construction. We consider three important factors of each trajectory to construct the preference dataset: (i) the correctness of the answer, (ii) the length, and (iii) the intrinsic confidence based on SC in Eq. (1). We emphasize that

SC is a complementary to external factors, which estimates trajectory-level fine-grained qualities and determines the strength of pairwise preferences.

Inspired by Shen et al. (2025a), we categorize preference pairs into two types: **Conciseness Pairs (CPs)**, formed by two correct trajectories where the preferred one is shorter; and **Deliberation Pairs (DPs)**, formed by two incorrect traces where the preferred one is longer. Unlike prior approaches that calibrate preference strength using per-question fixed budgets or external difficulty estimation, CAT uses only model-internal evidence to modulate pairwise preference score s .

For each input question x and its K candidate reasoning paths, we consider the margin between both lengths and self-certainty to acquire the preference score as s . Specifically, given a candidate pair (x, y_w, y_l) , we first compute the margin in terms of self-certainty, lengths, and correctness:

$$\begin{aligned} \Delta r &= r(y_w) - r(y_l) \\ \Delta SC &= SC(x, y_w) - SC(x, y_l) \end{aligned} \quad (2)$$

where $SC(\cdot)$ can be acquired by Eq.(1) and $r(\cdot)$ is a factor with respect to reasoning lengths and correctness:

$$r(y) = \begin{cases} +\frac{1}{|y|} & \text{if } y \text{ is correct} \\ -\frac{1}{|y|} & \text{if } y \text{ is incorrect} \end{cases} \quad (3)$$

This design assigns the highest reward to short, correct paths while imposing the lightest penalty on long, incorrect paths. Conversely, short but incorrect paths receive the most severe penalty.

For CPs, our intent is to favor short and confident solutions and reject long and unconfident ones. We therefore multiply r and SC so that a pair receives stronger strength precisely when the winning path y_w is not only much more efficient but also more internally decisive:

$$s_{CP}(x, y_w, y_l) = \Delta r \cdot \Delta SC \quad (4)$$

For DPs, we want to prefer long and unconfident attempts over short and confident failures, discouraging premature yet decisive mistakes. Accordingly, we reverse the confidence term, making wrong trajectories with larger certainty receive stronger penalties:

$$s_{DP}(x, y_w, y_l) = \Delta r \cdot (-\Delta SC) \quad (5)$$

In both cases, larger scores of preference pairs indicate potentially stronger and more discriminative preference signals for subsequent optimization.

Thus, we devise a **Dynamic Pruning** strategy to select the preference optimization dataset based on s . Concretely, for each query x , we rank the CP and DP sets by their score s in descending order, respectively, and retain only those pairs whose scores fall within the top three highest score levels. We then pool candidates from all the queries and sort them globally based on the preference score s , truncating the list by removing the bottom τ fraction, where τ denotes the truncation ratio. Finally, to prevent over-representing queries that produce many high-scoring pairs, we enforce a per-query cap and retain at most one CP and one DP per query in the final preference dataset.

3.2.2 Confidence-Weighted Preference Optimization (CWPO)

To adjust the model’s reasoning depth conditionally on its internal certainty, rather than applying a uniform length bias to all the samples, we propose the **Confidence-Weighted Preference Optimization (CWPO)** objective that pioneers the use of intrinsic self-certainty directly within the alignment loss landscape. Compared with the vanilla SimPO objective (Meng et al., 2024), we dynamically modulated the scaling factors of the winning and losing terms. Formally, the CWPO loss is computed as:

$$\mathcal{L}_{CWPO}(\pi_\theta) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(\frac{\beta_w}{|y_w|} \log \pi_\theta(y_w|x) - \frac{\beta_l}{|y_l|} \log \pi_\theta(y_l|x) - \gamma \right) \right] \quad (6)$$

where the dynamic weights β_w and β_l are acquired by the original β_{base} in SimPO and a **calibration ratio** (ρ) based on self-certainty and lengths:

$$\rho(x, y) = \frac{SC(x, y)}{|y|^\alpha} \quad (7)$$

where $\alpha \in (0, 1)$ is a length-aware exponent to keep SC and $|y|$ on a comparable scale for numerical stability. This ratio imposes an additional tunable length penalty so that confidence-guided scaling can better align gradient allocation with efficiency.

The CWPO loss sets different weights for conciseness pairs (CPs) and deliberation pairs (DPs): For CPs, where the model compares two correct responses, we define $\beta_w = \beta_{base} \cdot \sigma(\rho(x, y_w))$, while symmetrically scaling the loser’s weight using the inverse ratio $\beta_l = \beta_{base} \cdot \sigma(\rho(x, y_l)^{-1})$. This specifically incentivizes the model to commit to reasoning paths that are both correct and concise. Conversely, for DPs, we focus on penalizing short and

Model	MATH 500					AIME 2024					GPQA				
	Acc \uparrow	Len \downarrow	C-Len \downarrow	CR \uparrow	C-CR \uparrow	Acc \uparrow	Len \downarrow	C-Len \downarrow	CR \uparrow	C-CR \uparrow	Acc \uparrow	Len \downarrow	C-Len \downarrow	CR \uparrow	C-CR \uparrow
R1-1.5B	82.9	5087	3595	–	–	28.9	16609	8473	–	–	34.0	10100	9240	–	–
OverThink _{SimPO}	82.3	3899	2482	23.4%	31.0%	27.8	12104	6224	27.1%	26.6%	35.7	8908	7967	11.8%	13.8%
DAST	85.5	4189	2635	17.7%	26.7%	30.0	13911	7667	16.2%	9.5%	36.7	9980	8993	1.2%	2.7%
ConCISE _{SimPO}	80.8	3449	2300	32.2%	36.0%	25.6	10947	5663	34.1%	33.2%	34.9	7890	6995	21.9%	24.3%
CAT (Ours)	86.1	4087	2736	19.7%	23.9%	31.1	12619	7155	24.0%	15.6%	37.0	9716	8348	3.8%	9.7%
R1-7B	91.7	3657	3200	–	–	53.3	12831	7631	–	–	49.2	8553	7382	–	–
OverThink _{SimPO}	89.6	2421	1873	33.8%	41.5%	52.2	10030	5845	21.8%	23.4%	49.8	7270	5856	15.0%	20.7%
DAST	93.5	2997	2469	18.0%	22.8%	56.7	10804	7158	15.8%	6.2%	52.0	8068	7021	5.7%	4.9%
ConCISE _{SimPO}	89.5	2187	1805	40.2%	43.6%	48.9	8854	5333	31.0%	30.1%	50.2	6030	5011	29.5%	32.1%
CAT (Ours)	93.9	2431	1953	33.5%	39.0%	58.9	9880	5401	23.0%	29.2%	54.0	7781	6425	9.0%	13.0%
Qwen3-8B	96.2	5549	5088	–	–	74.4	14437	12060	–	–	59.9	7241	6354	–	–
OverThink _{SimPO}	94.9	3478	3107	37.3%	38.9%	73.3	10545	7960	27.0%	34.0%	57.1	3712	3389	48.7%	46.7%
DAST	96.5	4402	3964	20.7%	22.1%	75.6	13433	10252	7.0%	15.0%	59.9	6380	5523	11.9%	13.1%
ConCISE _{SimPO}	94.1	2880	2551	48.1%	49.9%	74.4	10190	7961	29.4%	34.0%	58.6	3939	3464	45.6%	45.5%
CAT (Ours)	96.6	3546	3205	36.1%	37.0%	76.7	11774	9866	18.4%	18.2%	60.9	4819	4393	33.4%	30.9%

Table 1: Accuracy (Acc), the mean response length over all trajectories (Len) and trajectories with correct final answers (C-Len), the percentage reduction in Len relative to the base model (CR), and the percentage reduction in C-Len relative to the base model (C-CR) on three benchmark datasets, respectively.

erroneous answers with unearned confidence. We set the penalty weight $\beta_l = \beta_{base} \cdot \sigma(\rho(x, y_l))$ and the winner’s weight $\beta_w = \beta_{base} \cdot \sigma(\rho(x, y_w)^{-1})$. By integrating these internal signals, CWPO moves beyond static length penalties, allowing the model to autonomously judge when to compress reasoning and when to deliberate, achieving a balance between efficiency and accuracy.

4 Experiments

4.1 Settings

Models and Datasets. We conduct comparative experiments on three LRMs: DeepSeek-R1-Distill-Qwen-7B (R1-7B) / 1.5B (R1-1.5B) (DeepSeek-AI, 2025) and Qwen3-8B (Yang et al., 2025). For the training dataset, following Qiao et al. (2025), we randomly select 2,000 questions from the MATH training set (Hendrycks et al., 2021), maintaining diversity in both difficulty and response length.

Benchmarks. We follow Shen et al. (2025a) to select three benchmarks, including MATH-500 (Lightman et al., 2024), AIME24 (MAA, 2024), and GPQA (Rein et al., 2023).

Baselines. We select several state-of-the-art methods for efficient reasoning as baselines, including OverThink (Chen et al., 2024), DAST (Shen et al., 2025a), and ConCISE (Qiao et al., 2025). For ConCISE, we choose the best-performing alternative ConCISE_{SimPO} as the comparison baseline. All of these methods are under the paradigm of preference

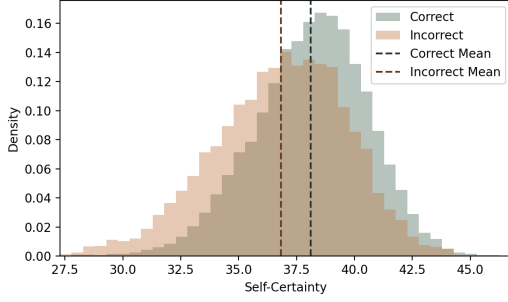
optimization with the SimPO objective.

Implementation Details. Following Shen et al. (2025a), we generate 20 candidate responses per question in our training set, and set the maximum sequence length to 4,096 tokens. Based on the hyperparameter analysis in Appendix B.2, the truncation ratio (τ) was set to 0.15. The preference optimization is conducted within the SimPO framework (Meng et al., 2024). We adopt low-rank adaptation (LoRA) (Hu et al., 2022) with the rank of $r = 32$, scaling factor of $\alpha = 64$, dropout rate of 0.05. The training epoch is 1 while the batch size is 16. The learning rate is $5e-5$ for DeepSeek-R1-Distill-Qwen-7B and Qwen3-8B, and $5e-6$ for DeepSeek-R1-Distill-Qwen-1.5B, as the weaker 1.5B backbone yields more DPs after CAPL and thus benefits from more conservative optimization. All the experiments are conducted on 2 NVIDIA A800 GPUs. More training details are provided in Appendix A.1. Decoding processes are executed using the the OpenR1 evaluation scripts (Hugging Face, 2025), with comprehensive decoding details provided in Appendix A.2. The experimental results are presented with mean values over 3 runs.

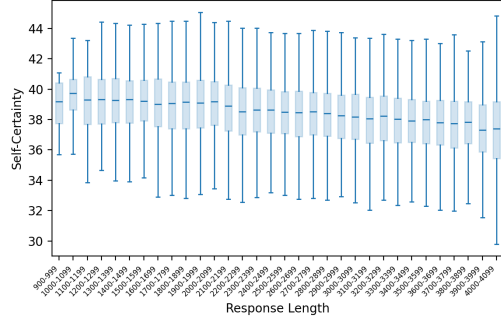
4.2 Results and Analysis

4.2.1 Overall Results

The results in Table 1 show that CAT achieves the highest accuracy (exceeding the backbone model) on all the three benchmarks while maintaining an



(a) The distribution of SC for correct and incorrect responses.



(b) Box plots of SC across varying response lengths.

Figure 2: Analysis of Self-Certainty (SC) distributions regarding response correctness and robustness to length on the MATH dataset (Level 4), derived from 20 reasoning paths per question generated by Qwen3-8B ($L_{max} = 4096$).

Benchmark	Method	Acc	Len	C-Len	CR	C-CR
MATH-500	Origin	91.7	3657	3200	–	–
	CAT	93.9	2431	1953	33.5%	39.0%
	w/o CWPO	93.6	2869	2361	21.5%	26.2%
	w/o CAPL	93.0	2512	2014	31.3%	37.1%
AIME24	Origin	53.3	12831	7631	–	–
	CAT	58.9	9880	5401	23.0%	29.2%
	w/o CWPO	53.3	10859	6387	15.4%	16.3%
	w/o CAPL	56.7	9131	5807	28.8%	23.9%
GPQA	Origin	49.2	8553	7382	–	–
	CAT	54.0	7781	6425	9.0%	13.0%
	w/o CWPO	51.5	7877	6343	7.9%	14.1%
	w/o CAPL	50.5	7745	6412	9.4%	13.1%

Table 2: Ablation study of Confidence-Aware Preference Labeling and Confidence-Weighted Preference Optimization on DeepSeek-R1-Distill-Qwen-7B.

acceptable compression rate (CR), suggesting that CAT can allocate reasoning steps adaptively to obtain better performance. Although OverThink and ConCISE attain the most substantial compression rates, they still incur an unavoidable loss in accuracy relative to the backbone model. DAST and CAT exhibit similar balancing trends between task performance and compression, as both aim to achieve adaptive compression while preserving the model’s reasoning capability. Compared with DAST, CAT delivers higher for all the base models, and achieves higher CR and C-CR in most settings. These results suggest that CAT is more effective at adaptive reasoning, demonstrating effectiveness of our proposed confidence-aware adaptive reasoning approach based on the model’s intrinsic signals.

4.2.2 Ablation Study

To assess the key components in CAT, including Confidence-Aware Preference Labeling (CAPL) and Confidence-Weighted Preference Optimization

(CWPO), we conduct an detailed ablation study by removing either CAPL (w/o CAPL, where preferences are scored only by Δr) or CWPO (w/o CWPO, where we replace CWPO with vanilla SimPO). The results in Table 2 show that all these parts contribute to the final performance. We observe that constructing preference pairs solely based on length differences (w/o CAPL) yields a higher compression ratio but leads to a larger degradation in reasoning performance on most tasks. This observation highlights the importance of high-quality training data that provides difficulty-adaptive reasoning signals.

Due to the paper limit, we further explore the effect of Self-Certainty on CAT in Appendix B.1.

4.2.3 Analysis of Self-Certainty

To better understand how self-certainty helps the model achieve an optimal balance between reasoning accuracy and length, we conduct a detailed analysis on the reasoning trajectories generated by Qwen3-8B of the MATH dataset.

SC effectively distinguishes correct and incorrect reasoning paths. In Figure 2a, we analyze the distribution of self-certainty for correct and incorrect responses on MATH (Level 4). The distributions for correct and incorrect responses concentrate around distinct means, with correct responses exhibiting a higher mean. This suggests that SC can effectively distinguish correct from incorrect reasoning trajectories and is strongly correlated with response quality.

SC is robust to reasoning lengths. We analyze self-certainty across responses of varying lengths. As illustrated in Figure 2b, SC is not noticeably affected by response lengths: across the entire length range, the median (blue line) shows only a very

Q: A triangle has sides measuring 1 unit and 3 units. The length of the third side is an integer value of units. What is the length of the third side, in units?

[Ground Truth: 3]

backbone model 981 tokens self-certainty: 17.1

[Thinking Process]

<think>
 ... Let me denote the third side as 'x'.
 According to the triangle inequality... $x < 4$... $x > 2$... So,
 combining the first two inequalities... $2 < x < 4$.
 ... The integers greater than 2 and less than 4 are 3.
 Wait, let me double-check that. If x is 3...
 ... $1 + 3 > 3$... true.
 ... $1 + 3 = 4 > 3$... true.
 ... $3 + 3 > 1$... true.
 ... Is there any other integer that could work? ... no, 3 is the
 only integer that works.
 Wait, just to be thorough, what if x was 2? ... No...
 What about x=1? ... No...
 How about x=4? ... No...
 ... there might be cases where the triangle is degenerate...
 Therefore, the only possible integer value for the third side
 is 3 units.

</think>

[Final Response]

... The only integer in this range is 3.
 Therefore, The length of the third side is 3.

DAST 656 tokens self-certainty: 16.6

[Thinking Process]

<think>
 ... The two given sides are 1 unit and 3 units, and
 the third side has to be an integer... Let me denote
 the third side as x...
 Let me list the three inequalities... $4 > x$, which
 means $x < 4$... $x > 2$... $x > -2$... this doesn't really
 add any new information... So... x has to be greater
 than 2 and less than 4.
 But wait, x is an integer... 3 is the only integer
 greater than 2 and less than 4.
 Wait, let me check if that works.
 Let me plug x=3 into the inequalities... So yes, x=3
 works.
 But wait, could there be another integer? ...
 integers are 3 only... So, the third side must be 3
 units long.

</think>

[Final Response]

... $2 < x < 4$... the only possible value is 3.
 Therefore, the length of the third side is 3 units.

CAT 573 tokens self-certainty: 17.8

[Thinking Process]

<think>
 ... So, the sides are 1, 3, and x, where x is an
 integer. Applying the triangle inequality...
 First inequality: $1 + 3 > x \rightarrow 4 > x \rightarrow x < 4$.
 Second inequality: $1 + x > 3 \rightarrow x > 2$.
 Third inequality: $3 + x > 1 \rightarrow x > -2$.
 ... So from the first two inequalities, x must be
 greater than 2 and less than 4... the possible
 integer values for x are 3.
 Wait, let me check that again... x can be 3 only
 because 3 is an integer between 2 and 4.
 Wait, but wait a second. If the two sides are 1 and
 3, and the third side is 3, does that form a valid
 triangle? ... So yes, all conditions are satisfied.
 So the only possible integer value for the third
 side is 3.

</think>

[Final Response]

The possible integer value for the third side is 3...
 Therefore, the length of the third side is: 3.

Figure 3: Case study on DeepSeek-R1-Distill-Qwen-7B. All the three methods solve the problem correctly. Compared with the backbone model and DAST that shorten the reasoning chain but lower self-certainty, CAT further reduces the reasoning length and yields higher self-certainty.

Benchmark	Method	Acc	Len	C-Len	CR	C-CR
MATH-500	Origin	91.7	3657	3200	-	-
	DPO	92.7	2827	2188	22.7%	31.6%
	CWPO _{DPO}	93.8	2722	2134	25.6%	33.3%
AIME24	Origin	53.3	12831	7631	-	-
	DPO	53.3	10187	6145	20.6%	19.5%
	CWPO _{DPO}	58.9	10720	5532	16.5%	27.5%
GPQA	Origin	49.2	8553	7382	-	-
	DPO	50.7	8353	7057	2.3%	4.4%
	CWPO _{DPO}	53.0	8326	6864	2.7%	7.0%

Table 3: Results of DPO and CWPO_{DPO} on DeepSeek-R1-Distill-Qwen-7B.

slight downward trend, which is largely attributable to the fact that shorter responses contain a higher proportion of correct trajectories, indicating that SC is stable regardless of reasoning lengths.

4.2.4 Generalization Across Preference Optimization Methods

To test the generalization ability of our method, we further apply our method to DPO in addition to SimPO, and assess the performance on DeepSeek-R1-Distill-Qwen-7B using the same preference pairs constructed by CAPL. The DPO-version CWPO objective (denoted as CWPO_{DPO}) is slightly different from vanilla CWPO, which is detailed in Appendix C. The results in Table 3 indicate that CWPO_{DPO} beats standard DPO in most of the metrics on three benchmarks, demonstrating the promising generalization ability to different preference optimization methods.

4.2.5 Case Study

To intuitively illustrate how CAT affects reasoning behaviors, we present a case study on DeepSeek-R1-Distill-Qwen-7B in Figure 3. We observe that all the three methods reach the correct answer but exhibit different reasoning lengths and self-certainty. The backbone model repeatedly verifies the same inequalities and explicitly checks invalid alternatives. DAST reduces the reasoning length with lower self-certainty, while still retaining additional verification beyond the core derivation. In comparison, CAT achieves the shortest reasoning chain with higher self-certainty. It does not eliminate reflection entirely: after deriving the feasible interval, it only keeps a brief validity check rather than exploring invalid candidates. This qualitative case supports the design of CAT, which incorporates self-certainty together with correctness and length signals to favor concise and confident correct reasoning paths.

5 Conclusion

This work proposes confidence-adaptive thinking (CAT), which addresses the pronounced overthinking and token overhead in large reasoning models through intrinsic confidence awareness. CAT integrates self-certainty as LRMs' intrinsic confidence to enable them to compress confident responses while deliberating on uncertain ones. Extensive experiments demonstrate that CAT consistently achieves a superior balance between inference efficiency and accuracy.

Limitations

While CAT demonstrates a superior balance between reasoning accuracy and efficiency, we identify the following areas for future improvement:

Path-Level Aggregation. Our current framework utilizes path-level Self-Certainty to score reasoning traces. While this metric effectively differentiates high-quality responses, aggregating token-level signals into a single scalar for the entire sequence may overlook variations in confidence at specific reasoning steps. Future work could explore integrating token positions with their specific Self-Certainty scores to enable more precise step-level compression.

Domain-Specific Evaluation. Our experiments focus on STEM disciplines such as mathematics and physics that allow rigorous correctness verification. Although Self-Certainty is an intrinsic signal independent of ground truth, our preference labeling strategy currently utilizes verification results. We aim to extend this approach to open-ended generation tasks where Self-Certainty can guide alignment without reliance on external answers.

Offline Optimization Paradigm. CAT employs Confidence-Weighted Preference Optimization on static datasets constructed from pre-sampled trajectories. This offline setting limits the ability of the policy to dynamically update its confidence estimates during the training process. Future research will investigate transitioning from offline optimization to online reinforcement learning variants, allowing the model to iteratively refine its reasoning efficiency through continuous interaction.

Acknowledgments

This work was supported by Sichuan Science and Technology Program (2025ZNSFSC1488), Non-communicable Chronic Diseases-National Science and Technology Major Project (2023ZD0501806), Fundamental Research Funds for the Central Universities (ZYGX2025XJ041), and CIPS-SMP-Zhipu Large Model Fund (CIPS-SMP20250314).

References

Pranjal Aggarwal and Sean Welleck. 2025. [L1: controlling how long A reasoning model thinks with reinforcement learning](#). *CoRR*, abs/2503.04697.

Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuqi

Liu, Mengfei Zhou, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. 2024. [Do NOT think that much for 2+3=? on the overthinking of o1-like llms](#). *CoRR*, abs/2412.21187.

Yu-Neng Chuang, Leisheng Yu, Guanchu Wang, Lizhe Zhang, Zirui Liu, Xuanting Cai, Yang Sui, Vladimir Braverman, and Xia Ben Hu. 2025. [Confident or seek stronger: Exploring uncertainty-based on-device LLM routing from benchmarking to generalization](#). *CoRR*, abs/2502.04428.

Yingqian Cui, Pengfei He, Jingying Zeng, Hui Liu, Xianfeng Tang, Zhenwei Dai, Yan Han, Chen Luo, Jing Huang, Zhen Li, Suhang Wang, Yue Xing, Jiliang Tang, and Qi He. 2025. [Stepwise perplexity-guided refinement for efficient chain-of-thought reasoning in large language models](#). In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 18581–18597. Association for Computational Linguistics.

DeepSeek-AI. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *CoRR*, abs/2501.12948.

Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, and Maxim Panov. 2024. [Fact-checking the output of large language models via token-level uncertainty quantification](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9367–9385, Bangkok, Thailand. Association for Computational Linguistics.

Sicheng Feng, Gongfan Fang, Xinyin Ma, and Xinchao Wang. 2025. [Efficient reasoning models: A survey](#). *Trans. Mach. Learn. Res.*, 2025.

Yichao Fu, Xuewei Wang, Yuandong Tian, and Jiawei Zhao. 2025. [Deep think with confidence](#). *CoRR*, abs/2508.15260.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koepl, Preslav Nakov, and Iryna Gurevych. 2024. [A survey of confidence estimation and calibration in large language models](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 6577–6595. Association for Computational Linguistics.

Tingxu Han, Zhenting Wang, Chunrong Fang, Shiyu Zhao, Shiqing Ma, and Zhenyu Chen. 2025. [Token-budget-aware LLM reasoning](#). In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, volume ACL 2025 of *Findings of ACL*, pages 24842–24855. Association for Computational Linguistics.

- Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. 2024. [Training large language models to reason in a continuous latent space](#). *CoRR*, abs/2412.06769.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. [Measuring mathematical problem solving with the MATH dataset](#). In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [Lora: Low-rank adaptation of large language models](#). In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.
- Hugging Face. 2025. [Open r1: A fully open reproduction of deepseek-r1](#).
- Zhewei Kang, Xuandong Zhao, and Dawn Song. 2025. [Scalable best-of-n selection for large language models via self-certainty](#). *CoRR*, abs/2502.18581.
- Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. [Let’s verify step by step](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Yue Liu, Jiaying Wu, Yufei He, Hongcheng Gao, Hongyu Chen, Baolong Bi, Jiaheng Zhang, Zhiqi Huang, and Bryan Hooi. 2025. [Efficient inference for large reasoning models: A survey](#). *CoRR*, abs/2503.23077.
- Haotian Luo, Li Shen, Haiying He, Yibo Wang, Shiwei Liu, Wei Li, Naiqiang Tan, Xiaochun Cao, and Dacheng Tao. 2025. [O1-pruner: Length-harmonizing fine-tuning for o1-like reasoning pruning](#). *ArXiv*, abs/2501.12570.
- Xinyin Ma, Guangnian Wan, Runpeng Yu, Gongfan Fang, and Xinchao Wang. 2025. [Cot-valve: Length-compressible chain-of-thought tuning](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 6025–6035. Association for Computational Linguistics.
- Codeforces MAA. 2024. American invitational mathematics examination-aime 2024.
- Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. [SimpO: Simple preference optimization with a reference-free reward](#). In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel J. Candès, and Tatsunori Hashimoto. 2025. [s1: Simple test-time scaling](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 20275–20321. Association for Computational Linguistics.
- Tergel Munkhbat, Namgyu Ho, Seo Hyun Kim, Yongjin Yang, Yujin Kim, and Se-Young Yun. 2025. [Self-training elicits concise reasoning in large language models](#). In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, volume ACL 2025 of *Findings of ACL*, pages 25127–25152. Association for Computational Linguistics.
- Sania Nayab, Giulio Rossolini, Giorgio C. Buttazzo, Nicolamaria Manes, and Fabrizio Giacomelli. 2024. [Concise thoughts: Impact of output length on LLM reasoning and cost](#). *CoRR*, abs/2407.19825.
- Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M. Waleed Kadous, and Ion Stoica. 2025. [Routellm: Learning to route llms from preference data](#). In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.
- OpenAI. 2024. [Openai o1 system card](#). *CoRR*, abs/2412.16720.
- Ziqing Qiao, Yongheng Deng, Jiali Zeng, Dong Wang, Lai Wei, Guanbo Wang, Fandong Meng, Jie Zhou, Ju Ren, and Yaoxue Zhang. 2025. [Concise: Confidence-guided compression in step-by-step efficient reasoning](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 8010–8029. Association for Computational Linguistics.
- Xiaoye Qu, Yafu Li, Zhaochen Su, Weigao Sun, Jianhao Yan, Dongrui Liu, Ganqu Cui, Daizong Liu, Shuxian Liang, Junxian He, Peng Li, Wei Wei, Jing Shao, Chaochao Lu, Yue Zhang, Xian-Sheng Hua, Bowen Zhou, and Yu Cheng. 2025. [A survey of efficient reasoning for large reasoning models: Language, modality, and beyond](#). *CoRR*, abs/2503.21614.
- David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. [GPQA: A graduate-level google-proof q&a benchmark](#). *CoRR*, abs/2311.12022.
- Matthew Renze and Erhan Guven. 2024. [The benefits of a concise chain of thought on problem-solving in large language models](#). In *2nd International Conference on Foundation and Large Language Models, FLLM 2024, Dubai, United Arab Emirates, November 26-29, 2024*, pages 476–483. IEEE.

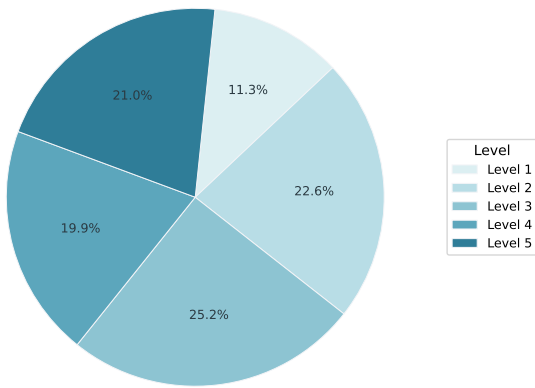
- Yi Shen, Jian Zhang, Jieyun Huang, Shuming Shi, Wenjing Zhang, Jiangze Yan, Ning Wang, Kai Wang, Zhaoxiang Liu, and Shiguo Lian. 2025a. [DAST: difficulty-adaptive slow-thinking for large reasoning models](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025 - Industry Track, Suzhou, China, November 4-9, 2025*, pages 2322–2331. Association for Computational Linguistics.
- Zhenyi Shen, Hanqi Yan, Linhai Zhang, Zhanghao Hu, Yali Du, and Yulan He. 2025b. [CODI: compressing chain-of-thought into continuous space via self-distillation](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 677–693. Association for Computational Linguistics.
- Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, Na Zou, Hanjie Chen, and Xia Hu. 2025. [Stop overthinking: A survey on efficient reasoning for large language models](#). *Trans. Mach. Learn. Res.*, 2025.
- Hanshi Sun, Momin Haider, Ruiqi Zhang, Huitao Yang, Jiahao Qiu, Ming Yin, Mengdi Wang, Peter L. Bartlett, and Andrea Zanette. 2024. [Fast best-of-n decoding via speculative rejection](#). In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Yue Wang, Qiuzhi Liu, Jiahao Xu, Tian Liang, Xingyu Chen, Zhiwei He, Linfeng Song, Dian Yu, Juntao Li, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. 2025. [Thoughts are all over the place: On the underthinking of o1-like llms](#). *CoRR*, abs/2501.18585.
- Yuyang Wu, Yifei Wang, Tianqi Du, Stefanie Jegelka, and Yisen Wang. 2025. [When more is less: Understanding chain-of-thought length in llms](#). *CoRR*, abs/2502.07266.
- Heming Xia, Chak Tou Leong, Wenjie Wang, Yongqi Li, and Wenjie Li. 2025. [TokenSkip: Controllable chain-of-thought compression in LLMs](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3351–3363, Suzhou, China. Association for Computational Linguistics.
- Fengli Xu, Qianyu Hao, Chenyang Shao, Zefang Zong, Yu Li, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, Yuwei Yan, Qinglong Yang, Yiwen Song, Sijian Ren, Xinyuan Hu, Jie Feng, Chen Gao, and Yong Li. 2025. [Toward large reasoning models: A survey of reinforced reasoning with large language models](#). *Patterns*, 6(10):101370.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 40 others. 2025. [Qwen3 technical report](#). *CoRR*, abs/2505.09388.
- Qiyong Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, and 16 others. 2025. [DAPO: an open-source LLM reinforcement learning system at scale](#). *CoRR*, abs/2503.14476.
- Zhiyuan Zeng, Qinyuan Cheng, Zhangyue Yin, Bo Wang, Shimin Li, Yunhua Zhou, Qipeng Guo, Xuanjing Huang, and Xipeng Qiu. 2024. [Scaling of search and learning: A roadmap to reproduce o1 from reinforcement learning perspective](#). *CoRR*, abs/2412.14135.
- Jiaxin Zhang. 2025. [Confidence-aware reasoning: Optimizing self-guided thinking trajectories in large reasoning models](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025 - Industry Track*, pages 2081–2095. Association for Computational Linguistics.

A Experimental Details

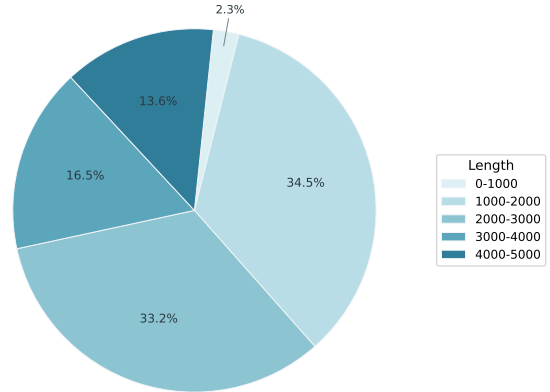
A.1 Training details

Training Set. While the construction methodology of the training set is detailed in Section 4, Figures 4a and 4b illustrate its diversity in terms of difficulty and length. Specifically, the difficulty distribution aligns with the *Level* metric of the 2000 selected questions from the MATH dataset. The length distribution reflects the reasoning chains generated by DeepSeek-R1-Distill-Qwen-7B, obtained by sampling 20 paths per question with a temperature of 0.6 and a top_p of 0.95.

Training Configurations. Following Dynamic Pruning with a truncation ratio of $\tau = 0.15$, our constructed preference pairs for R1-7B yielded 1,765 Conciseness Pairs (CPs; 93.1%) and 130 Deliberation Pairs (DPs; 6.9%), a distribution similar to that reported in DAST (Shen et al., 2025a). Similarly, for R1-1.5B, we identified 1,742 CPs (88.8%) and 221 DPs (11.2%). For Qwen3-8B, the resulting dataset comprised 1,517 CPs (97.4%) and 40 DPs (2.6%). These results indicate that for models with stronger reasoning capabilities, such as Qwen3-8B, Confidence-Aware Preference Labeling generates a higher proportion of CPs to facilitate the learning of conciseness. Conversely, for models with weaker reasoning abilities, such as R1-1.5B, the method



(a) Difficulty diversity of training set.



(b) Length diversity of training set.

Figure 4: Difficulty and length distributions illustrating the diversity of the question set.

Benchmark	Method	Acc	Len	C-Len	CR	C-CR
MATH-500	Origin	91.7	3657	3200	–	–
	CAT	93.9	2431	1953	33.5%	39.0%
	w/o SC in CWPO only	93.2	2771	2218	24.2%	30.7%
	w/o SC in CAPL & CWPO	92.6	2511	1859	31.3%	41.9%
AIME24	Origin	53.3	12831	7631	–	–
	CAT	58.9	9880	5401	23.0%	29.2%
	w/o SC in CWPO only	53.3	9088	6237	29.2%	18.3%
	w/o SC in CAPL & CWPO	50.0	10507	4906	18.1%	35.7%
GPQA	Origin	49.2	8553	7382	–	–
	CAT	54.0	7781	6425	9.0%	13.0%
	w/o SC in CWPO only	51.0	8036	6489	6.0%	12.1%
	w/o SC in CAPL & CWPO	49.5	7561	6185	11.6%	16.2%

Table 4: Ablation study of Self-Certainty (SC) on DeepSeek-R1-Distill-Qwen-7B.

produces more DPs to encourage cautious exploration. Furthermore, following DAST, the original SimPO hyperparameters were set to $\beta = 200$ and $\gamma = 1$ for all three models. All baselines use comparable data budgets to CAT.

A.2 Evaluation Details

In our evaluation setup, we use a unified decoding configuration for all experiments, with temperature = 0.6 and top_p = 0.95 (DeepSeek-AI, 2025). The maximum generation length is capped at 32,768 tokens for all the models.

B Additional Experiments

B.1 Ablation Study on Self-Certainty

To further investigate the effects of Self-Certainty (SC) in CAT, we conduct an additional ablation study to remove SC in CWPO or CAPL & CWPO, respectively. Note that removing SC in CAPL only is equivalent to w/o CAPL in Table 2, which is not repeatedly analyzed in this section.

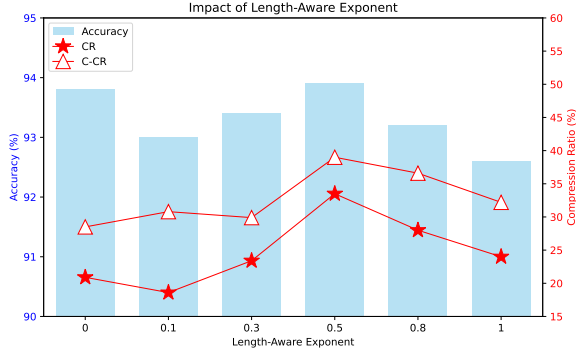
Table 4 shows that SC matters in both data construction and model optimization. Although removing SC commonly brings a higher compression rate, the reasoning accuracy largely degrades on all the benchmarks. This directly supports that SC contributes non-redundant value to our method.

B.2 Analysis of Hyperparameters

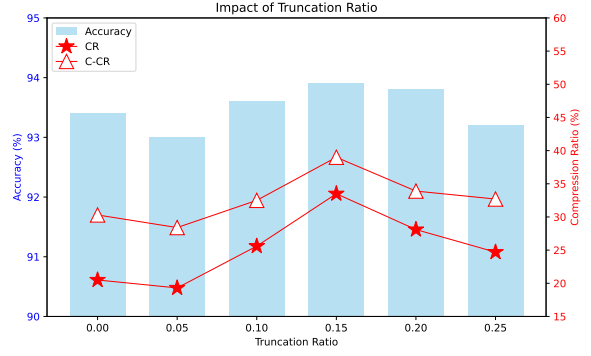
To examine the effects of the Length-Aware Exponent(α) and the Truncation Ratio (τ), we performed a grid search on MATH-500 using the R1-7B model for validation. As shown in Figure 5, R1-7B achieves the optimal results when $\alpha=0.5$ and $\tau=0.15$. In both single-parameter sweeps, these settings yield the best Acc and compression ratio. For the other backbone models, we conducted the same hyperparameter search and selected the final hyperparameters accordingly.

B.3 Validation of SC as a Difficulty Indicator

Because MATH Level 4 does not provide a sufficiently diverse difficulty distribution, we analyze SC across different difficulty levels on the training set. As shown in Figure 7, as the difficulty level increases, the average SC over all trajectories (including both correct and incorrect ones) exhibits a clear downward trend. This suggests that, beyond separating correct from incorrect trajectories, SC also captures differences in problem difficulty. This property is central to CAT: it enables difficulty-aware control.

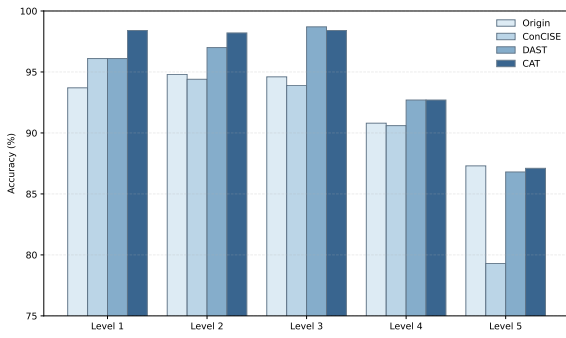


(a) Impact of varying the Length-Aware Exponent α on model performance.

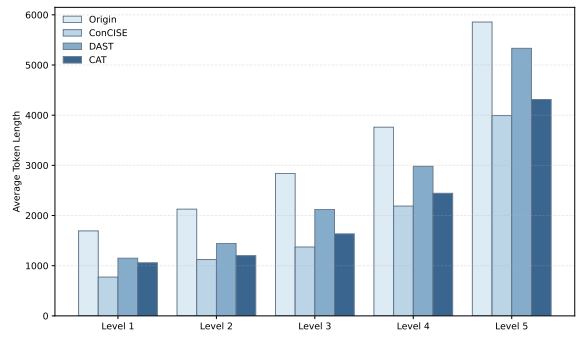


(b) Impact of varying the Truncation Ratio τ on model performance.

Figure 5: Hyperparameter analysis.



(a) Accuracy Performance on Difficulty Levels.



(b) Average Token Length on Difficulty Levels.

Figure 6: Performance comparison on MATH-500 across different difficulty levels.

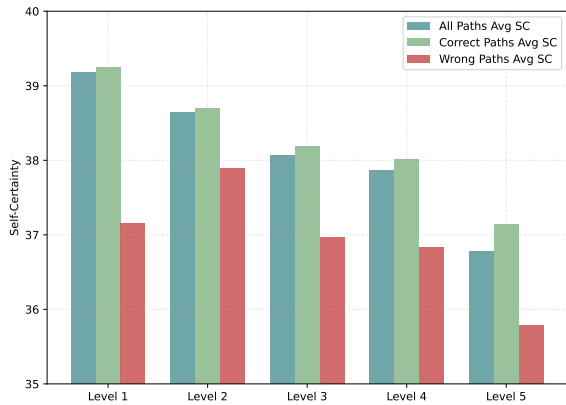


Figure 7: Average Self-Certainty (SC) across difficulty levels on the training set, derived from 20 reasoning paths per question generated by Qwen3-8B. A clear downward trend in SC is evident for all, correct, and incorrect trajectories as the difficulty level increases from Level 1 to Level 5. This indicates that SC effectively captures problem difficulty beyond correctness, serving as a reliable signal for difficulty-aware control.

B.4 Performance Analysis Across Difficulty Levels

We analyze the performance of CAT, DAST, and CONCISE on the MATH-500 dataset using the R1-7B model across varying difficulty levels. As shown in Figure 6, CAT achieves the highest accuracy in the two most challenging difficulty tiers, demonstrating a substantial advantage over other baselines and highlighting its robust reasoning capabilities for complex problems. While DAST also exhibits difficulty adaptability, it underperforms CAT in both accuracy and length compression. Furthermore, although CONCISE achieves the most significant length reduction, its reasoning performance deteriorates sharply as problem difficulty increases, creating a marked gap compared to the other methods and indicating a lack of capability in handling complex reasoning tasks.

C Details of CWPO_{DPO} Objective

To examine whether the benefits of CWPO-style SC weighting can generalize beyond the original CWPO setting, we incorporate the same SC-based

dynamic weighting strategy into the DPO objective. The resulting objective, denoted as CWPO_{DPO} , is formulated as follows:

$$\mathcal{L}_{\text{CWPO}_{\text{DPO}}}(\pi_{\theta}; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(\beta_w \log \frac{\pi_{\theta}(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta_l \log \frac{\pi_{\theta}(y_l|x)}{\pi_{\text{ref}}(y_l|x)} \right) \right] \quad (8)$$

Here, we use the same definitions of β_w and β_l as in Section 3.2.2.