

Temporalizing Confidence: Evaluation of Chain-of-Thought Reasoning with Signal Temporal Logic

Zhenjiang Mao¹, Artem Bisliouk^{1,2}, Rohith Reddy Nama¹, Ivan Ruchkin¹

University of Florida¹, University of Mannheim²

{z.mao,a.bisliouk,namarohithreddy,iruchkin}@ufl.edu

Abstract

Large Language Models (LLMs) have shown impressive performance in mathematical reasoning tasks when guided by Chain-of-Thought (CoT) prompting. However, they tend to produce highly confident yet incorrect outputs, which poses significant risks in domains like education, where users may lack the expertise to assess reasoning steps. To address this, we propose a structured framework that models stepwise confidence as a temporal signal and evaluates it using Signal Temporal Logic (STL). In particular, we define formal STL-based constraints to capture desirable temporal properties and compute robustness scores that serve as structured, interpretable confidence estimates. Our approach also introduces a set of uncertainty reshaping strategies to enforce smoothness, monotonicity, and causal consistency across the reasoning trajectory. Experiments show that our approach consistently improves calibration metrics and provides more reliable uncertainty estimates than conventional confidence aggregation and post-hoc calibration.

1 Introduction

Large language models (LLMs) are increasingly applied in educational contexts such as concept explanation, question answering, and personalized tutoring, especially in STEM domains like mathematics (Kasneji et al., 2023). These models exhibit strong capabilities in solving complex problems; however, they also tend to produce answers that are fluent and seemingly confident, yet factually incorrect. In educational settings, such outputs can be particularly problematic, as students may lack the expertise to distinguish between correct and incorrect reasoning (Polyxeni Paulina Kastania, 2024), and may be misled by responses that appear trustworthy. This mismatch between confidence and correctness raises critical concerns about the reliability of LLM-generated answers and highlights

the importance of integrating uncertainty estimation into educational AI systems.

Although prior work has explored various uncertainty estimation techniques, such as predictive entropy, sampling-based variance, and confidence calibration, most studies focus on general NLP tasks rather than educational scenarios (Zhao et al., 2021; Jiang et al., 2021). In educational contexts like mathematics learning, well-calibrated uncertainty can be especially useful for guiding student attention, supporting teacher oversight, and improving feedback systems. However, existing uncertainty metrics often show poor alignment with actual correctness (Zhu et al., 2025).

In this work, we address this challenge by proposing a novel approach to *estimate uncertainty in LLM-based chain-of-thought (CoT) reasoning* for high school mathematics problems. Our method models the sequence of reasoning steps as a temporal confidence signal and evaluates its structural properties using Signal Temporal Logic (STL) (Fainekos and Pappas, 2006). Instead of modifying the LLM or directly penalizing its outputs, we quantify undesirable confidence behaviors, such as abrupt increases following uncertain steps, by computing robustness scores against formal STL constraints. This yields a constraint-aware aggregation scheme that captures how confidence is expected to evolve over time, offering a structured and interpretable view of the reasoning process while improving calibration. We evaluate our method on a curated dataset of Chinese Gaokao mathematics multiple-choice questions (Zhang et al., 2023). Experimental results show that our method significantly improves calibration, reducing Expected Calibration Error (ECE) compared to baseline uncertainty aggregation methods.

This paper’s contributions are: (1) a novel perspective that treats stepwise confidence in chain-of-thought reasoning as a temporal signal amenable to formal analysis, (2) a constraint-aware modeling

approach that reshapes confidence trajectories and quantifies their structural quality using STL robustness and (3) empirical validation of our method’s effectiveness on Chinese Gaokao mathematics. Fig 1 provides an overview of our pipeline: starting from stepwise CoT confidence, we apply uncertainty reshaping followed by STL-based temporal logic evaluation. This transformation results in interpretable, structure-aware confidence scores.

2 Related Work

Applications of LLMs in Education: LLMs have been widely adopted in educational settings for tasks such as grammar correction, content generation, problem explanation, and intelligent tutoring. Prior studies highlight their potential to support learners across domains like language writing (Kasneji et al., 2023), mathematics (Gan et al., 2023), and personalized feedback (Zhou et al., 2025; Wang et al., 2024). For instance, LLMs like MathGPT and Khanmigo have been used to generate step-by-step math explanations aligned with curriculum standards (Shah et al., 2024), while ChatGPT has shown promise in automated feedback for student essays and short answers (Kasneji et al., 2023). Despite these advances, concerns remain around academic integrity, hallucinated outputs, and students over-relying on unverified responses (Benítez et al., 2024). Moreover, few works explicitly quantify how reliable these educational outputs are, or how uncertainty signals can be used to guide learners or inform teachers. As LLMs become integral to education technology, recent surveys have called for deeper investigations into trust, transparency, and uncertainty in educational applications (Idris et al., 2024).

Uncertainty in LLMs: LLMs exhibit remarkable fluency across diverse NLP tasks, yet their outputs often suffer from overconfidence and miscalibration (Kadavath et al., 2022), especially in impact-sensitive domains such as education. Existing work has explored various uncertainty estimation techniques, including entropy-based methods like predictive entropy and confidence gaps (Zhu et al., 2025), as well as sampling consistency across outputs (Lyu et al., 2025). Confidence calibration is another active area, revealing that LLMs tend to be overconfident, particularly in zero-shot or out-of-domain tasks (Desai and Durrett, 2020; Zhao et al., 2021). Recent studies also propose using uncertainty signals to guide reasoning, such as

Uncertainty-Guided CoT prompting (Zhu et al., 2025), active prompting for data selection, and consistency-based calibration (Diao et al., 2023; Lyu et al., 2025). However, most of these methods have been evaluated on general NLP or code generation tasks, with limited attention to structured educational settings like math problem solving, where reliable uncertainty estimates can help students assess model-generated reasoning and assist teachers in diagnosing student understanding..

CoT and STL: Recent advances in CoT prompting have significantly improved the multi-step reasoning ability of LLMs, yet they also introduce new layers of uncertainty, such as error propagation across intermediate steps and unfaithful explanations (Zhang et al., 2022; Wang et al., 2022; Tanneru et al., 2024). In this work, we propose a novel perspective that treats CoT steps as discrete temporal signals, enabling the use of STL to formally specify and evaluate reasoning quality over time (Rescher and Urquhart, 2012). STL allows for expressive specifications like *eventually correct* or *always consistent*, and provides quantitative robustness scores that capture the degree of satisfaction or violation (Fainekos and Pappas, 2006). This formalism has been successfully applied in domains such as motion planning (van Huijgevoort et al., 2024), reinforcement learning (Li et al., 2017), and control synthesis, and offers a promising path toward interpretable and rigorous evaluation of LLM-generated reasoning trajectories. Applying STL to CoT would not only enable structured detection of flawed reasoning patterns but also facilitate the development of confidence-aware feedback and scoring systems in education and other applications.

3 STL-Guided Confidence Estimation

Our approach consists of three stages: (1) generating a stepwise confidence signal via CoT prompting, (2) applying uncertainty reshaping strategies to promote temporal consistency, and (3) evaluating the reshaped sequence using STL. This pipeline is illustrated in Fig 1(b), showing how each reshaping strategy transforms a sample confidence trajectory. Compared to the original signal, our smoothing strategies effectively suppress abrupt spikes and produce a more temporally coherent confidence trajectory, while preserving the overall trend of reasoning. This behavior is crucial for downstream STL-based evaluation, which benefits

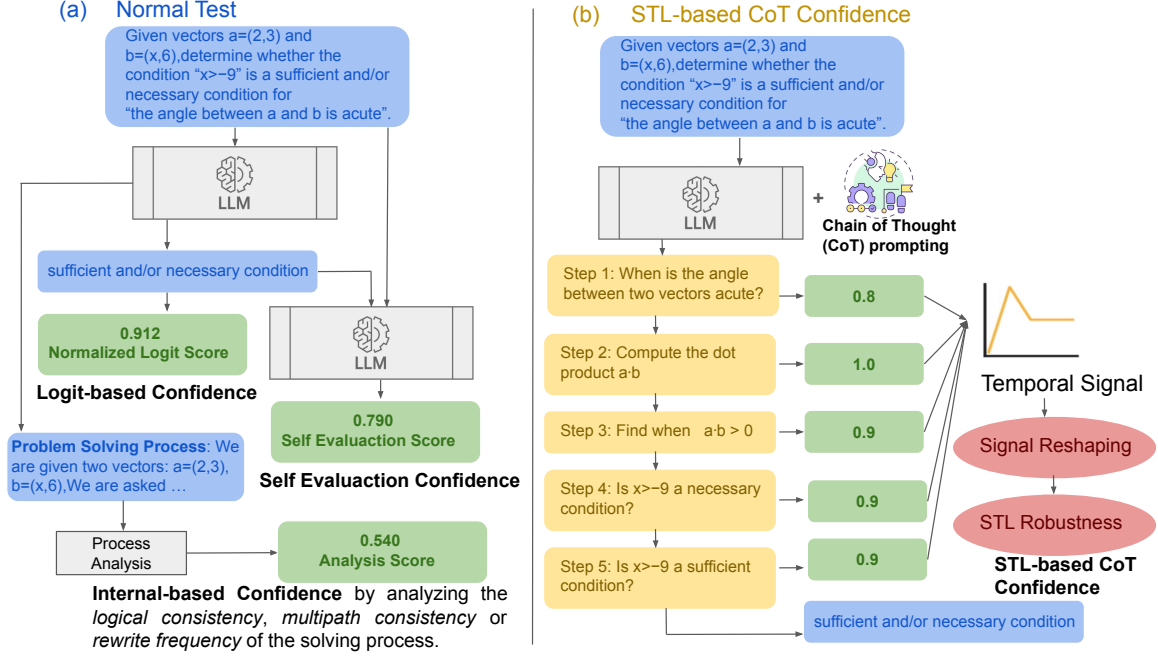


Figure 1: (a) Conventional methods output global confidence via logit, self-evaluation, or internal analysis. (b) Our method models step-wise CoT confidence as a temporal signal, applies signal reshaping, and evaluates robustness using STL to obtain a temporally consistent confidence score.

from smoother and more causally consistent input signals.

3.1 Problem Setup

We model LLMs as autonomous agents tasked with solving high school mathematics problems. Given an input question $q \in \mathcal{Q}$, the agent generates a final answer $a \in \mathcal{A}$ along with a scalar uncertainty score $u \in [0, 1]$, representing its confidence in the answer. Ideally, high confidence should correspond to high correctness probability, and vice versa (Guo et al., 2017).

To evaluate calibration, we employ the *Expected Calibration Error (ECE)* (Naeni et al., 2015; De-sai and Durrett, 2020), a widely-used metric that quantifies the mismatch between confidence and accuracy. Formally, we partition predictions into M bins based on their confidence values and compute:

$$\text{ECE} = \sum_{m=1}^M \frac{|B_m|}{n} |\text{acc}(B_m) - \text{conf}(B_m)|, \quad (1)$$

where B_m denotes the set of predictions falling into the m -th confidence bin, $\text{acc}(B_m)$ is the empirical accuracy defined as

$$\text{acc}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \mathbf{1}[\hat{a}_i = a_i], \quad (2)$$

and $\text{conf}(B_m)$ is the average predicted confidence. Here, n is the total number of examples, and $\mathbf{1}[\cdot]$ is the indicator function that returns 1 if the prediction is correct, and 0 otherwise. Since the task is formulated as multiple-choice classification, both model predictions and ground-truth answers are represented as one of a finite set of discrete options (e.g., A, B, C, D). This allows correctness to be determined via exact match of the selected option label, avoiding ambiguities arising from natural language variation. rect , and 0 otherwise.

In addition to ECE, we also report the *Brier Score (BS)* as a complementary calibration metric. The Brier Score measures the mean squared error between predicted confidence and ground-truth correctness, defined as:

$$\text{BS} = \frac{1}{n} \sum_{i=1}^n (c_i - y_i)^2, \quad (3)$$

where $c_i \in [0, 1]$ is the model’s predicted confidence for example i , and $y_i \in \{0, 1\}$ is the binary correctness label. Lower Brier Scores indicate better calibrated and more reliable confidence estimates.

Unlike conventional classification tasks, reasoning in LLMs often unfolds over multiple steps (Wei et al., 2022). This raises an additional challenge: confidence should not only be calibrated across

Question	Logit	Self-Eval	Internal
Given vectors $\mathbf{a} = (2, 3)$ and $\mathbf{b} = (x, 6)$. Determine whether the condition “ $x > -9$ ” is a sufficient and/or necessary condition for “the angle between \mathbf{a} and \mathbf{b} is acute.” Options: A. Sufficient but not necessary B. Necessary but not sufficient C. Sufficient and necessary D. Neither sufficient nor necessary <i>Model answer: C (incorrect), True answer: B</i>	0.99	0.98	0.99
Set parabola $C : y^2 = 4x$. The focus is F . A line passes through $(-2, 0)$ with slope $\frac{2}{3}$, intersecting C at points M and N . Compute $F\vec{M} \cdot F\vec{N}$. Options: A. 5 B. 6 C. 7 D. 8 (correct) <i>Model answer: D (correct)</i>	0.98	0.95	0.97

Table 1: Examples of high school mathematics questions and confidence scores from three estimation strategies: Logit-based, Self-evaluation, and Internal consistency. While the model is highly confident in all three views, the first question is incorrectly answered, leading to significant miscalibration. The **Expected Calibration Error (ECE)** for Logit, Self-Eval, and Internal confidences are **0.485**, **0.465**, and **0.480**, respectively.

examples, but also evolve smoothly and consistently over the reasoning trajectory (Zhu et al., 2025). Hence, the problem extends to generating temporally coherent uncertainty sequences that reflect both local confidence (per step) and global correctness (final answer). Our objective is thus to design a framework where uncertainty estimates are not only well-calibrated across examples, but also evolve in a temporally consistent manner—exhibiting properties such as smooth progression, causal coherence, and alignment with the underlying reasoning process.

3.2 Uncertainty Reshaping Strategies

To model reasoning-time uncertainty, we use CoT prompting to elicit a sequence of intermediate reasoning steps $\{s_1, \dots, s_T\}$, each associated with a confidence score $c_t \in [0, 1]$. We treat the resulting confidence sequence $\mathbf{c} = \{c_1, \dots, c_T\}$ as a temporal signal (Rescher and Urquhart, 2012). However, due to the inherent causal nature of reasoning, abrupt increases in confidence, especially after initially low-confidence steps, can be misleading (Zhu et al., 2025). To take advantage of this insight, we propose several signal reshaping functions that induce smoother and causally consistent confidence evolution. While some of these strategies are conceptually related to smoothing methods in time-series analysis, they are, to the best of our knowledge, novel in the context of modeling stepwise confidence in LLM-based reasoning.

- **Causal Minimum Smoothing (CMS):** Limits future confidence based on past minimum values plus a small fixed margin δ :

$$\tilde{c}_t = \min \left(c_t, \min_{i < t} c_i + \delta \right)$$

- **Exponential Decay Smoothing (EDS):** Applies exponential smoothing by blending the current value with the average of past values:

$$\tilde{c}_t = \alpha \cdot c_t + (1 - \alpha) \cdot \frac{1}{t} \sum_{i=1}^{t-1} c_i$$

- **Monotonic Penalty Smoothing (MPS):** Dampens confidence spikes if the previous step is below a fixed threshold τ . This focuses on upward spikes, which are more likely to mislead following uncertain steps:

$$\tilde{c}_t = \begin{cases} \frac{c_{t-1} + c_t}{2} & \text{if } c_{t-1} < \tau \text{ and } c_t > c_{t-1} \\ c_t & \text{otherwise} \end{cases}$$

- **Guarded Smoothing (GS):** Caps sudden jumps beyond threshold τ plus tolerance ϵ :

$$\tilde{c}_t = \begin{cases} \tau + \epsilon, & \text{if } c_{t-1} < \tau \text{ and } c_t > \tau + \epsilon \\ c_t, & \text{otherwise} \end{cases}$$

The reshaped sequence $\tilde{\mathbf{c}} = \{\tilde{c}_1, \dots, \tilde{c}_T\}$ is passed to a formal temporal logic evaluation module described next as shown in Fig. 1(b).

3.3 STL-Based Temporal Evaluation

Rather than relying solely on the final-step confidence c_T or averaging all stepwise confidences, we propose a STL-based framework to evaluate the temporal structure of the confidence trajectory $\tilde{\mathbf{c}} = \{\tilde{c}_1, \dots, \tilde{c}_T\}$ (Fainekos and Pappas, 2006). STL enables formal specification of desired temporal properties of confidence during reasoning, such as smooth progression or eventual certainty.

Each STL formula encodes a specific temporal pattern, and its associated robustness score

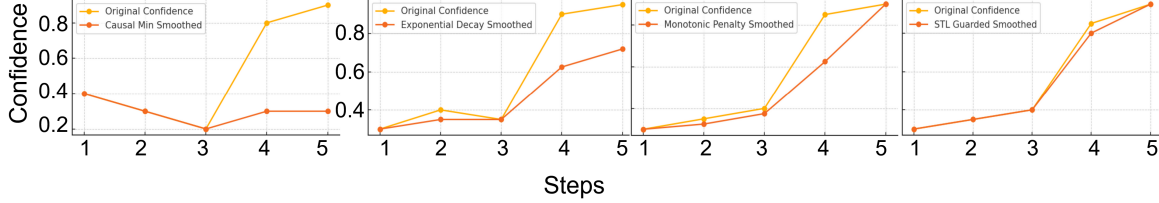


Figure 2: Visualization examples of stepwise confidence signals before and after applying different uncertainty reshaping strategies. Each subplot compares the original confidence trajectory (yellow) against a smoothed version (orange) using one of the following methods: (left to right) Causal Minimum Smoothing (CMS), Exponential Decay Smoothing (EDS), Monotonic Penalty Smoothing (MPS), and Guarded Smoothing (GS). These transformations produce smoother and more temporally consistent confidence profiles while preserving the overall trend.

$\rho_i = \rho(\tilde{c}, \text{STL}) \in \mathbf{R}$ quantifies how well the reshaped signal satisfies that property (Donzé and Maler, 2010). A positive score indicates the satisfaction margin, while a negative score represents the magnitude of a violation.

We define three STL specifications, each yielding a separate confidence score:

- **Eventually Confident:** Confidence should eventually rise above a threshold τ :

$$\text{STL1} = \diamond_{[t_1, t_2]}(\tilde{c}(t) > \tau)$$

- **Always Stable or Increasing:** Confidence should not drop abruptly:

$$\begin{aligned} \Delta \tilde{c}(t) &= \tilde{c}(t) - \tilde{c}(t-1) \\ \text{STL2} &= \square_{[t_1, t_2]}(\Delta \tilde{c}(t) \geq -\epsilon) \end{aligned}$$

- **Local Smoothness:** Confidence should not change too much between steps:

$$\text{STL3} = \square_{[t_1, t_2]}(|\Delta \tilde{c}(t)| \leq \delta)$$

Each resulting score $\hat{c} = \text{ReLU}(\rho(\tilde{c}, \text{STL})) \in [0, 1]$ represents an interpretable, temporally-informed confidence score derived from logic-based robustness. These scores can be used independently for analysis or combined in multi-dimensional calibration evaluation.

While our STL-based scoring framework does not require labeled data during reasoning, it does rely on threshold hyperparameters (e.g., τ, ϵ, δ) that influence robustness computation. To set them, we perform a grid search on a held-out validation set. This makes our approach partially *post-hoc* in nature: only the STL evaluation stage requires data-driven tuning, whereas the preceding confidence reshaping is fully unsupervised and model-agnostic.

4 Experiments

In Section 4, we present an ablation study and a comparison against established post-hoc calibration techniques that investigates the impact of STL parameterization and compares our method against established post-hoc calibration techniques such as *Temperature Scaling* (Guo et al., 2017) and *Histogram Binning* (Zadrozny and Elkan, 2001). Our results show that STL-based evaluation provides not only competitive calibration performance but also interpretable, temporally grounded diagnostics of reasoning quality.

Experimental Setup: We conduct our experiments using the Qwen-7B language model,¹ a high-performing open-source LLM optimized for Chinese and mathematical reasoning tasks. Our evaluation is conducted on all multiple-choice questions from Chinese national college entrance exams (Gaokao) spanning 2010 to 2022, totaling 12 years of official high school mathematics problems. These questions are drawn from GAOKAO-Bench (Zhang et al., 2023), to assess LLMs’ language understanding and symbolic reasoning capabilities using real-world exam data. Table 1 illustrates representative examples.

To prevent the model from relying on surface-level pattern matching or memorized templates, we augment the dataset following strategies inspired by GSM-Symbolic (Mirzadeh et al., 2024), which shows that LLMs often fail when symbols, numbers, or phrasing are changed. We use a more advanced reasoning model – OpenAI’s o1 API to perform all paraphrasing operations in a controlled and semantically faithful manner. Specifically, each original problem is rewritten to preserve logical structure and correct answer while varying lexical expressions (e.g., transforming “find

¹<https://huggingface.co/Qwen/Qwen-7B>

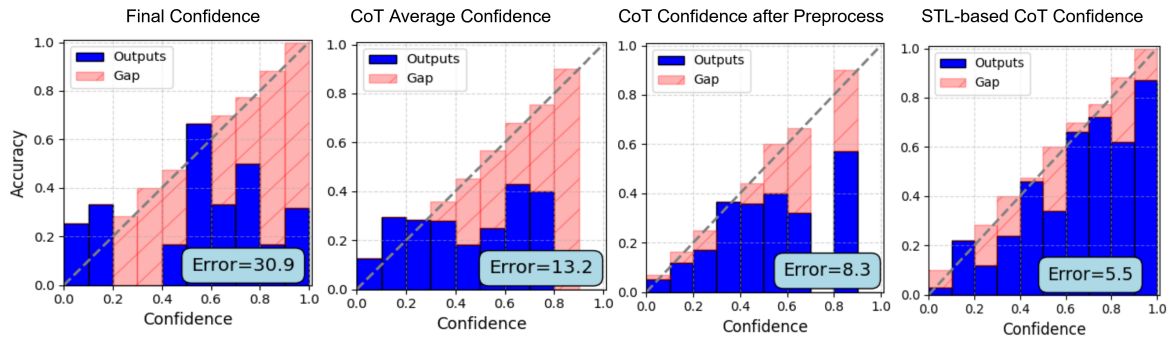


Figure 3: ECE comparison of four confidence estimation methods. (1) Final-step confidence; (2) average confidence over CoT steps; (3) CoT confidence after applying Uncertainty Reshaping Strategies; and (4) STL-based CoT confidence, which combines Uncertainty Reshaping Strategies with the **STL1** formula (*Eventually Confident*: confidence should eventually exceed a threshold τ). The STL1-enhanced method achieves the best calibration (ECE = 5.5).

the intersection of sets A and B” into “determine the elements shared by both A and B”). We further introduce linguistic variation through backtranslation, translating problems into a pivot language (such as French) and then back to English, thereby injecting natural noise without changing semantics. Additionally, symbolic formulations are diversified using template-based transformations—for example, the set expression “ $A \cap B$ ” may be rephrased as “ $x \in Z, \sqrt{x} \leq 4$ ” or “ $B = \{x \in Z \mid x^2 \leq 16\}$ ”. These augmentations collectively evaluate the model’s robustness to paraphrasing, symbol rewriting, and structural variation, ensuring assessment focuses on genuine reasoning rather than memorized syntax. The correct answer for each augmented sample is inherited from the corresponding original Gaokao-Bench problem, since paraphrasing and symbolic transformations preserve semantic and logical equivalence. In total, the original 432 questions are each rewritten twice, resulting in a final dataset of 1,296 problem instances for evaluation.

Quantitative Results and Analysis: To illustrate how different estimation methods affect calibration behavior, Figure 3 shows an example confidence histogram under four representative strategies. While it only reflects a single problem instance and one STL constraint (STL1), the figure demonstrates how reshaping and temporal logic evaluation yield more aligned and interpretable confidence estimates. We now turn to aggregate results across the full test set. All scores are averaged over three runs, and \pm denotes standard deviation caused by the randomness introduced through

temperature-controlled decoding, which affects the variability and creativity of LLM outputs. Table 2 presents the Expected Calibration Error (ECE) for various estimation strategies across three types of uncertainty sources: logits-based, self-evaluation-based, and internal-based. Table 3 complements this with Brier Scores, which jointly capture calibration and sharpness of probabilistic estimates.

From the ECE results, we observe that traditional post-hoc calibration methods such as Temperature Scaling (Guo et al., 2017) and Histogram Binning (Zadrozny and Elkan, 2001) reduce miscalibration compared to raw one-step uncertainty. For example, Histogram Binning achieves an ECE of 0.139 on logits-based predictions, improving substantially over the one-step baseline (0.324). However, these methods operate globally and do not account for the multi-step nature of reasoning in LLMs.

In contrast, CoT-based methods yield stronger performance, especially when combined with our proposed Uncertainty Reshaping Strategies. Simply averaging confidence across CoT steps reduces ECE across all sources, and applying smoothing techniques such as Causal Minimum Smoothing (CMS) or Exponential Decay Smoothing (EDS) brings further gains. For instance, CMS reduces logits-based ECE to 0.107, the lowest among all non-STL methods.

STL-based temporal evaluation further improves calibration. By enforcing high-level temporal constraints like *Eventually Confident* (STL1), *Always Stable* (STL2), and *Locally Smooth* (STL3), the model’s confidence trajectory becomes more inter-

Method	Reshaping Strategy	Logits-based ↓	Self-evaluation-based ↓	Internal-based ↓
1-step Uncertainty	-	0.324 ± 0.045	0.692 ± 0.035	0.694 ± 0.033
Temperature Scaling	-	0.246 ± 0.061	0.158 ± 0.033	0.173 ± 0.046
Histogram Binning	-	0.139 ± 0.004	0.095 ± 0.069	0.185 ± 0.129
CoT Average	-	0.141 ± 0.062	0.542 ± 0.039	0.603 ± 0.037
	CMS	0.107 ± 0.048	0.486 ± 0.040	0.579 ± 0.041
	EDS	0.126 ± 0.022	0.502 ± 0.036	0.573 ± 0.039
	MPS	0.129 ± 0.063	0.530 ± 0.037	0.602 ± 0.038
	GS	0.140 ± 0.060	0.538 ± 0.038	0.579 ± 0.021
STL1 (Eventually Confident)	-	0.174 ± 0.019	0.082 ± 0.021	0.102 ± 0.055
	CMS	0.250 ± 0.198	0.136 ± 0.006	0.119 ± 0.023
	EDS	0.236 ± 0.064	0.098 ± 0.012	0.500 ± 0.497
	MPS	0.211 ± 0.019	0.080 ± 0.017	0.100 ± 0.046
	GS	0.212 ± 0.026	0.077 ± 0.018	0.096 ± 0.035
STL2 (Always Stable)	-	0.170 ± 0.071	0.113 ± 0.011	0.075 ± 0.040
	CMS	0.153 ± 0.021	0.126 ± 0.013	0.074 ± 0.030
	EDS	0.114 ± 0.063	0.114 ± 0.015	0.056 ± 0.028
	MPS	0.188 ± 0.009	0.118 ± 0.015	0.063 ± 0.013
	GS	0.164 ± 0.074	0.111 ± 0.013	0.070 ± 0.039
STL3 (Locally Smooth)	-	0.184 ± 0.021	0.154 ± 0.015	0.099 ± 0.031
	CMS	0.122 ± 0.018	0.081 ± 0.037	0.084 ± 0.030
	EDS	0.149 ± 0.008	0.076 ± 0.040	0.083 ± 0.034
	MPS	0.171 ± 0.025	0.151 ± 0.030	0.083 ± 0.046
	GS	0.180 ± 0.017	0.118 ± 0.006	0.091 ± 0.043

Table 2: Expected Calibration Error (ECE) comparison across confidence sources and estimation strategies. STL-based methods, particularly STL1–STL3 combined with Uncertainty Reshaping (CMS, EDS), consistently yield better calibration than traditional techniques.

pretable and aligned with reasoning quality. STL1 combined with GS achieves an ECE of 0.077 on self-evaluation-based confidence and 0.096 on internal-based, outperforming all other approaches. Notably, STL2 with EDS reaches an ECE of 0.056 on internal-based confidence—the best result across all settings.

The Brier Score analysis mirrors this trend. STL methods consistently produce lower scores than CoT average or standard post-hoc techniques. STL1 with CMS achieves the best self-evaluation-based Brier Score (0.234), while STL2 with EDS offers the best internal-based score (0.056). These results confirm that applying STL logic not only improves calibration error but also leads to sharper, more reliable probability estimates.

Overall, STL-based confidence estimation outperforms traditional calibration and CoT-only baselines, particularly when paired with CMS or EDS. These findings highlight the value of structured temporal logic as a calibration framework for LLM-based reasoning, offering both theoretical guarantees and empirical gains.

5 Conclusion

This paper presents a structured approach to confidence estimation for LLM-based mathematical reasoning. By modeling stepwise confidence as a tem-

poral signal and evaluating its quality using STL, our method addresses limitations in traditional calibration and step-level aggregation techniques. We introduce a suite of uncertainty reshaping strategies and STL-based robustness constraints that enforce desirable properties such as eventual certainty, monotonic progression, and local smoothness.

Experimental results on GAOKAO-Bench demonstrate that our STL-based evaluation, particularly when paired with smoothing strategies like CMS and EDS, consistently achieves lower ECE and Brier Scores compared to standard baselines. Beyond quantitative improvements, the framework provides a principled, interpretable method for diagnosing and enhancing reasoning quality in educational LLM applications.

6 Limitations

While our method improves calibration and interpretability, it is currently limited to high school-level multiple-choice math problems. Extending the framework to open-ended questions, formal proofs, or multi-modal reasoning is a promising direction. Since all experiments are conducted on Qwen-7B, generalization to other models remains uncertain. Testing on models like Gemma 3, Llama 3.2, or DeepSeek would help assess robustness across architectures.

Method	Reshaping Strategy	Logits-based ↓	Self-evaluation-based ↓	Internal-based ↓
1-step Uncertainty	-	0.339 ± 0.019	0.677 ± 0.032	0.678 ± 0.032
Temperature Scaling	-	0.263 ± 0.040	0.255 ± 0.018	0.276 ± 0.018
Histogram Binning	-	0.284 ± 0.002	0.263 ± 0.020	0.259 ± 0.021
CoT Average	-	0.225 ± 0.033	0.498 ± 0.023	0.564 ± 0.028
	CMS	0.218 ± 0.032	0.442 ± 0.019	0.537 ± 0.029
	EDS	0.219 ± 0.032	0.455 ± 0.017	0.530 ± 0.027
	MPS	0.219 ± 0.033	0.483 ± 0.022	0.563 ± 0.028
	GS	0.222 ± 0.033	0.493 ± 0.022	0.536 ± 0.037
STL1 (Eventually Confident)	-	0.223 ± 0.003	0.244 ± 0.001	0.246 ± 0.001
	CMS	0.218 ± 0.005	0.234 ± 0.000	0.240 ± 0.000
	EDS	0.219 ± 0.003	0.238 ± 0.001	0.241 ± 0.001
	MPS	0.221 ± 0.000	0.243 ± 0.002	0.246 ± 0.001
	GS	0.223 ± 0.004	0.243 ± 0.002	0.239 ± 0.001
STL2 (Always Stable)	-	0.238 ± 0.008	0.252 ± 0.005	0.254 ± 0.002
	CMS	0.239 ± 0.009	0.246 ± 0.002	0.249 ± 0.001
	EDS	0.243 ± 0.004	0.253 ± 0.003	0.253 ± 0.001
	MPS	0.239 ± 0.006	0.252 ± 0.004	0.254 ± 0.002
	GS	0.238 ± 0.008	0.252 ± 0.004	0.257 ± 0.005
STL3 (Locally Smooth)	-	0.232 ± 0.011	0.236 ± 0.002	0.240 ± 0.001
	CMS	0.237 ± 0.008	0.241 ± 0.002	0.242 ± 0.001
	EDS	0.239 ± 0.006	0.242 ± 0.001	0.244 ± 0.000
	MPS	0.236 ± 0.010	0.236 ± 0.003	0.238 ± 0.002
	GS	0.233 ± 0.011	0.235 ± 0.001	0.249 ± 0.002

Table 3: Brier Score comparison across methods. Lower scores indicate better-calibrated and sharper probabilistic predictions. STL-based temporal constraints further improve performance beyond CoT averaging and post-hoc calibration.

Both the reshaping strategies and STL specifications are manually defined. Future work could explore learning them dynamically via reinforcement learning or differentiable logic, enabling more adaptive and data-driven calibration beyond manual tuning.

This paper focuses on linear chain-of-thought reasoning. More complex prompting paradigms, such as tree-of-thought (Yao et al., 2023), introduce branching and cyclic structures that pose new challenges for temporal modeling. Extending our evaluation to computation tree logic (CTL) to accommodate such structures would broaden its applicability to richer and more realistic cognitive processes.

Finally, our method is post-hoc and does not influence model inference. Integrating uncertainty feedback into real-time tutoring systems could enable dynamic intervention, hinting, and early error detection.

References

- Trista M Benítez, Yueyuan Xu, J Donald Boudreau, Alfred Wei Chieh Kow, Fernando Bello, Le Van Phuoc, Xiaofei Wang, Xiaodong Sun, Gilberto Ka-Kit Leung, Yanyan Lan, et al. 2024. Harnessing the potential of large language models in medical education: promise and pitfalls. *Journal of the American Medical Informatics Association*, 31(3):776–783.
- Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. *arXiv preprint arXiv:2003.07892*.
- Shizhe Diao, Pengcheng Wang, Yong Lin, Rui Pan, Xiang Liu, and Tong Zhang. 2023. Active prompting with chain-of-thought for large language models. *arXiv preprint arXiv:2302.12246*.
- Alexandre Donzé and Oded Maler. 2010. Robust satisfaction of temporal logic over real-valued signals. In *International Conference on Formal Modeling and Analysis of Timed Systems*, pages 92–106. Springer.
- Georgios E Fainekos and George J Pappas. 2006. Robustness of temporal logic specifications. In *International Workshop on Formal Approaches to Software Testing*, pages 178–192. Springer.
- Wensheng Gan, Zhenlian Qi, Jiayang Wu, and Jerry Chun-Wei Lin. 2023. Large language models in education: Vision and opportunities. In *2023 IEEE international conference on big data (BigData)*, pages 4776–4785. IEEE.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.
- Mohamed Diab Idris, Xiaohua Feng, and Vladimir Dyo. 2024. Revolutionising higher education: Unleashing the potential of large language models for strategic transformation. *IEEE Access*.
- Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language

- models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.
- Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.
- Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günemann, Eyke Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.
- Xiao Li, Cristian-Ioan Vasile, and Calin Belta. 2017. Reinforcement learning with temporal logic rewards. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3834–3839. IEEE.
- Qing Lyu, Kumar Shridhar, Chaitanya Malaviya, Li Zhang, Yanai Elazar, Niket Tandon, Marianna Apidianaki, Mrinmaya Sachan, and Chris Callison-Burch. 2025. Calibrating large language models with sample consistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 19260–19268.
- Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. 2024. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. *arXiv preprint arXiv:2410.05229*.
- Mahdi Pakdaman Naeni, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.
- Nikoleta Polyxeni Paulina Kastania. 2024. Building trust in ai education: Addressing transparency and ensuring. *Trust and Inclusion in AI-mediated Education: Where Human Learning Meets Learning Machines*, page 73.
- Nicholas Rescher and Alasdair Urquhart. 2012. *Temporal logic*, volume 3. Springer Science & Business Media.
- Vedant Shah, Dingli Yu, Kaifeng Lyu, Simon Park, Jiatong Yu, Yinghui He, Nan Rosemary Ke, Michael Mozer, Yoshua Bengio, Sanjeev Arora, et al. 2024. Ai-assisted generation of difficult math questions. *arXiv preprint arXiv:2407.21009*.
- Sree Harsha Tanneru, Chirag Agarwal, and Himabindu Lakkaraju. 2024. Quantifying uncertainty in natural language explanations of large language models. In *International Conference on Artificial Intelligence and Statistics*, pages 1072–1080. PMLR.
- Birgit C van Huijgevoort, Ruohan Wang, Sadegh Soudjani, and Sofie Haesaert. 2024. Specification-guided temporal logic control for stochastic systems: a multi-layered approach. *arXiv preprint arXiv:2407.03896*.
- Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S Yu, and Qingsong Wen. 2024. Large language models for education: A survey and outlook. *arXiv preprint arXiv:2403.18105*.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.
- Bianca Zadrozny and Charles Elkan. 2001. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Icml*, volume 1.
- Xiaotian Zhang, Chunyang Li, Yi Zong, Zhengyu Ying, Liang He, and Xipeng Qiu. 2023. Evaluating the performance of large language models on gaokao benchmark. *arXiv preprint arXiv:2305.12474*.
- Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*.
- Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, pages 12697–12706. PMLR.
- Yizhou Zhou, Mengqiao Zhang, Yuan-Hao Jiang, Xinyu Gao, Naijie Liu, and Bo Jiang. 2025. A study on educational data analysis and personalized feedback report generation based on tags and chatgpt. *arXiv preprint arXiv:2501.06819*.
- Yuqi Zhu, Ge Li, Xue Jiang, Jia Li, Hong Mei, Zhi Jin, and Yihong Dong. 2025. Uncertainty-guided chain-of-thought for code generation with llms. *arXiv preprint arXiv:2503.15341*.