EVIDENCE INFORMATION GAIN FOR LLM CONFIDENCE ESTIMATION: A DIFFUSION DECISION MODEL-INSPIRED APPROACH

Anonymous authors
Paper under double-blind review

ABSTRACT

Estimating the confidence of Large Language Models (LLMs) remains a challenging problem, particularly when models generate complex reasoning processes. A prominent approach is to move beyond assigning a single confidence score to the entire reasoning chain and instead estimate trajectory-level confidence, capturing how confidence evolves throughout the reasoning process. In neuroscience, the Diffusion Decision Model (DDM) models decision-making as sequential evidence accumulation under noise, where a decision is reached when accumulated evidence crosses a decision boundary. Inspired by this perspective, we view the reasoning steps generated by an LLM as sequential evidence accumulated toward its final answer, and hypothesize that the contribution of reasoning steps to the final decision is informative for confidence estimation. To operationalize this idea, we first introduce a framework for quantifying the information gain of each evidence with respect to the final answer. We then derive novel features from the resulting evidence-accumulation patterns and use them to estimate the confidence of the model's final prediction. Across seven diverse benchmarks, our approach consistently outperforms existing confidence estimation methods, achieving meaningful improvements in AUROC while reducing ECE. Importantly, our method is applicable in both white-box and black-box settings, with the black-box variant achieving confidence estimation performance very close to that of the white-box setting.

1 INTRODUCTION

As LLMs are increasingly deployed in knowledge-intensive and decision-making applications, reliably assessing the confidence of their outputs is essential for trustworthy use. Confidence estimation provides a quantitative measure of the likelihood that an answer is correct, enabling systems to identify reliable predictions and flag uncertain outputs for verification. Prior work has shown that LLMs can provide meaningful estimates of their own correctness, highlighting the potential of confidence as a key signal for model reliability and calibration (Kadavath et al., 2022; Jiang et al., 2021). Thus, accurate confidence estimation is an important component of building trustworthy and dependable LLM-based systems.

Despite its importance, reliably estimating confidence in LLMs remains challenging, as model confidence does not always align with the correctness of their outputs. In particular, verbalized confidence can be poorly calibrated and may reflect the model's expressed commitment rather than its actual likelihood of being correct (Xiong et al., 2024; Kadavath et al., 2022). Moreover, the reliability of confidence estimates can vary substantially across models, tasks, and reasoning settings (Geng et al., 2024). Although substantial efforts have been devoted to improving confidence estimation and calibration in LLMs (Geng et al., 2024), further improvements are still needed to obtain confidence signals that more faithfully reflect answer correctness.

Existing approaches to LLM confidence estimation derive confidence from predictive probabilities, verbalized confidence, generation agreement, or internal representations (Geng et al., 2024). Most of these methods summarize the model's output or generation into a single confidence estimate, whereas *Trajectory-Level Confidence Estimation* explicitly considers how confidence evolves

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



Figure 1: **Evidence-Guided Confidence Estimation from Reasoning Dynamics.** The reasoning chain is divided into sequential evidence segments, from which trajectory-level features are extracted using information gain under two complementary measures based on **loss** and **self-consistency**, followed by a logistic regression model for confidence estimation. **(A)** A GSM8K example in which the model produces the correct answer; the proposed method assigns high confidence, reflecting the progressively increasing agreement and informative evidence along the reasoning trajectory. **(B)** A GPQA example in which the model produces an incorrect answer; in contrast, the evidence dynamics yield substantially lower confidence, despite the model generating a coherent reasoning chain. **(C)** The same incorrect GPQA example as in (B), evaluated using a conventional token-probability-based baseline. While the arithmetic mean of token probabilities assigns a high confidence score (0.8597) to the incorrect answer, our method assigns a substantially lower confidence score (0.1853), highlighting the difference between token-level fluency and evidence-based confidence.

throughout the reasoning trajectory. This perspective can capture intermediate signals accumulated during Chain-of-Thought reasoning and provide a more fine-grained assessment of confidence in the final answer. Recent studies have begun to explore uncertainty dynamics along reasoning trajectories (Mao et al., 2025; 2026), highlighting the potential of reasoning trajectories as a source of informative confidence signals.

Inspired by the *Diffusion Decision Model (DDM)*, a neuroscience model of decision-making in which the drift rate characterizes the strength and direction of evidence accumulation, we view an LLM's reasoning trajectory as a sequential accumulation of evidence toward a final answer. Building on this perspective, we propose a novel information-theoretic approach to estimate the *information gain* of individual evidence items from changes in predictive loss and self-consistency-based agreement probabilities. These evidence-level estimates form a trajectory-level feature representation that captures how informative each reasoning step is for the final answer, enabling *Trajectory-Level Confidence* estimation. We then use a lightweight logistic regression model to predict the probability that the final answer is correct. As illustrated in Figure 1, our approach can assign substantially lower confidence to the incorrect answer in Example B than probability-based confidence measures, which assign a high score despite the incorrect final answer.

We evaluate the proposed approach across seven diverse benchmarks and multiple LLMs. The results demonstrate meaningful improvements in AUROC over existing baselines, while simultaneously reducing the ECE, indicating more accurate and better-calibrated confidence estimates. Our contributions can be summarized as follows:

- We propose a *Trajectory-Level Confidence Estimation* framework that derives confidence from the evolution of evidence throughout the reasoning process and can be applied to LLMs in a black-box setting, without requiring access to their internal representations or parameters, while maintaining confidence estimation performance very close to that of the white-box setting.
- We introduce an information-theoretic metric for quantifying the *information gain* contributed by individual evidence items in an LLM reasoning trajectory.
- We demonstrate consistent improvements in confidence estimation, with modest but meaningful gains in AUROC and reductions in ECE across seven diverse benchmarks and multiple LLMs.

2 Related Work

Foundations of Confidence Estimation. Reliable confidence estimation is crucial for assessing the trustworthiness of LLM outputs, particularly when plausible responses may be incorrect. A common class of approaches derives confidence from the model's predictive distribution, using signals such as token probabilities, entropy, and sequence-level likelihood or cumulative log-probability Ulmer et al. (2024). *Self-consistency* uses agreement across multiple reasoning paths as a confidence signal (Wang et al., 2023), while *semantic entropy* measures uncertainty over semantically distinct answers rather than surface forms (Kuhn et al., 2023).

Confidence Estimation for LLM Reasoning. Recent work has increasingly studied confidence estimation to improve the reliability and efficiency of LLM reasoning. CER Razghandi et al. (2025) incorporates confidence into self-consistency by weighting sampled reasoning paths according to their estimated reliability. *Learning When to Sample* Xiong et al. (2026) uses features extracted from reasoning trajectories to estimate confidence and adaptively allocate additional sampling to uncertain instances. *Diagnosing Multi-step Reasoning Failures* Liu et al. (2026) estimates confidence at individual reasoning steps using an Information Bottleneck formulation, enabling the localization of potentially erroneous reasoning. *Beyond the Last Answer* Hammoud et al. (2025) instead evaluates intermediate subthoughts through sampled continuations, aggregating intermediate predictions to improve answer reliability.

Reasoning-Based Signals for Confidence Estimation Recent work increasingly leverages reasoning trajectories and model-internal signals for confidence estimation. Reasoning-trace length has been identified as a simple uncertainty indicator Devic et al. (2025), while internal representations can encode information about reasoning correctness Chen et al. (2026). Other approaches exploit token-level confidence signals to filter unreliable reasoning trajectories Fu et al. (2026) or improve the reliability of verbalized confidence through extended reasoning Podolak & Verma (2025). Similarly, token-level predictive distributions have been used to derive self-certainty for best-of-N selection Kang et al. (2025). Collectively, these studies demonstrate that reasoning trajectories provide informative signals for assessing answer reliability, motivating methods that explicitly quantify how evidence accumulates throughout the reasoning process.

Trajectory-Level Confidence Estimation Recent work has moved beyond a single confidence score by modeling how uncertainty evolves throughout the reasoning process. *Tracing Uncertainty in Language Model "Reasoning"* Grünefeld et al. (2026) estimates token-level uncertainty with respect to both the next token and the final answer across the reasoning trace, and summarizes the resulting uncertainty trajectory using statistical features that can be used by lightweight predictors such as logistic regression. Similarly, *Temporalizing Confidence* Mao et al. (2025) represents stepwise confidence as a temporal signal and uses Signal Temporal Logic (STL) to characterize patterns such as smoothness, monotonicity, and causal consistency in the confidence trajectory. *Confidence over Time* Mao et al. (2026) further mines discriminative STL formulas from stepwise confidence signals to distinguish correct and incorrect reasoning, and uses hypernetworks to adapt their numerical parameters to individual questions. These studies highlight that collapsing the reasoning process into a single final confidence value can discard informative temporal structure, motivating approaches that explicitly model confidence dynamics throughout CoT reasoning.

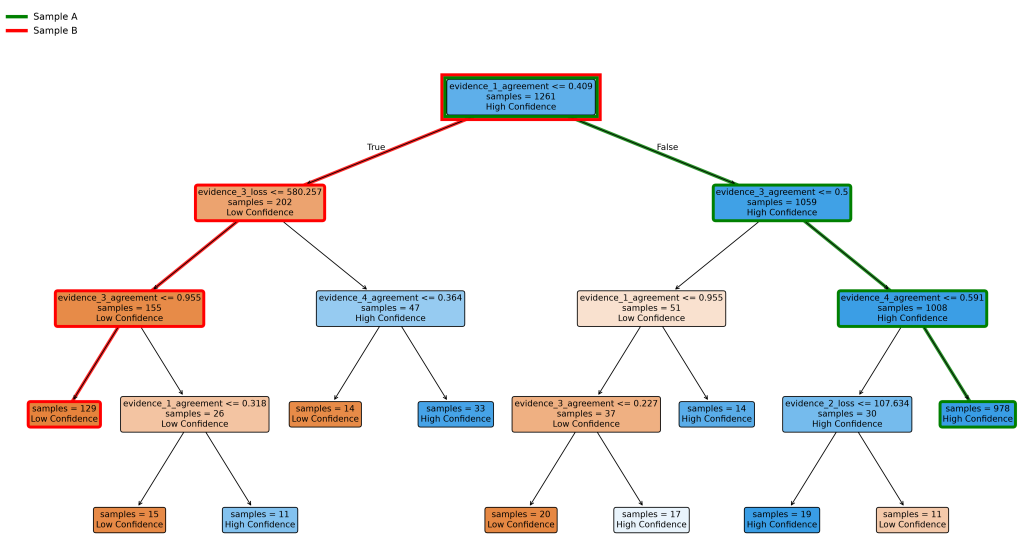


Figure 2: Decision tree visualization for analyzing feature contributions to confidence estimation. The tree is trained using the features of **CoT-EIG-Mean** with $K = 5$ evidences and is visualized up to depth four. The two highlighted examples illustrate how the learned feature-based decision rules distinguish between high- and low-confidence predictions: (a) a GSM8K example correctly classified as high confidence and (b) a GPQA example classified as low confidence.

3 METHOD

3.1 DIFFUSION DECISION MODEL

The Diffusion Decision Model (DDM) (Ratcliff, 1978; Ratcliff & McKoon, 2008) models decision making as sequential evidence accumulation under noise, where a decision is reached when accumulated evidence crosses a decision boundary. A key parameter, the *drift rate* (v), captures the rate and direction of evidence accumulation and reflects the quality and discriminativeness of the available information (Ratcliff & Tuerlinckx, 2002; Ratcliff et al., 2016). Thus, stronger evidence leads to faster accumulation toward the corresponding decision, whereas weak or ambiguous evidence produces slower and less consistent accumulation.

In addition to drift rate, the DDM includes the *boundary separation* (a), which represents the amount of evidence required before committing to a decision, and the *starting point* (z), which captures an initial bias toward one of the alternatives (Ratcliff & Tuerlinckx, 2002; Ratcliff & McKoon, 2008). The model can therefore distinguish the quality of decision evidence from response caution and decision bias (Ratcliff & McKoon, 2008; Ratcliff et al., 2016).

Overall, the DDM provides a principled framework for interpreting a decision not merely as an outcome, but as the result of an underlying evidence-accumulation process (Ratcliff et al., 2016). This perspective is particularly relevant when intermediate evidence contributing to a decision is available, as it provides a principled basis for relating the quality of such evidence to the strength of the resulting decision process.

3.2 QUANTIFYING EVIDENCE QUALITY VIA INFORMATION GAIN

In this work, we develop a DDM-inspired metric to quantify how strongly a chain-of-thought (CoT) reasoning trace influences a LLM's final decision. In the DDM framework, decision-making is viewed as an evidence-accumulation process, where *drift rate* measures the quality of information driving the decision process. Analogously, a LLM first generates a reasoning trace and then produces a final answer based on the accumulated information. We therefore treat the intermediate reasoning

states as evidence contributing to the final decision and quantify their influence by measuring how the accumulated reasoning information affects the final answer.

Given a prompt q, the LLM generates a reasoning trace R followed by a final answer Y, i.e., $q \xrightarrow{\text{LLM}} R\|Y$, where $\|$ denotes concatenation. We partition the reasoning trace into k sequential evidence segments, $R = e_1 \oplus \cdots \oplus e_k$, where $\oplus$ denotes concatenation of evidence segments, and define C_i as the reasoning prefix after observing the first i evidence segments: $C_i = C_{i-1} \oplus e_i$, with $C_0 = \emptyset$. Thus, C_{i-1} represents the reasoning state immediately before observing evidence e_i, while C_i represents the updated state after incorporating it. The quality of each evidence e_i, which we refer to as its *drift rate*, is quantified by its conditional mutual information with the final answer, as defined in Eq. 1.

$$\mathcal{Q}(e_i) \equiv I\left(Y; C_i \mid C_{i-1}\right), \tag{1}$$

where $I(Y; C_i \mid C_{i-1})$ measures the additional information that C_i provides about the final answer Y beyond C_{i-1}. Thus, $\mathcal{Q}(e_i)$ quantifies the incremental, decision-relevant contribution of evidence e_i, with larger values indicating greater additional information and smaller values indicating limited contribution beyond the preceding reasoning context. To compute $\mathcal{Q}(e_i)$, we use Equation 2; the detailed steps leading from the preceding formulation to this equation are provided in the Appendix A.1.

$$\mathbb{E}[\log p(Y \mid C_i)] - \mathbb{E}[\log p(Y \mid C_{i-1})] = I(Y; C_i \mid C_{i-1}). \tag{2}$$

Computing $I(Y; C_i \mid C_{i-1})$ requires expensive marginalization over the LLM distribution. To make the measure tractable, we replace the final answer Y with $Y_{>i}$, the continuation of the reasoning chain from e_i through the final answer. Although this substitution does not imply that $I(Y_{>i}; C_i \mid C_{i-1})$ is identical to $I(Y; C_i \mid C_{i-1})$, it preserves the underlying confidence-estimation objective: both formulations ultimately assess the reliability of arriving at the same final answer Y. Accordingly, Equation (2) is replaced by Equation 3).

$$\mathbb{E}[\log p(Y_{>i} \mid C_i)] - \mathbb{E}[\log p(Y_{>i} \mid C_{i-1})] = I(Y_{>i}; C_i \mid C_{i-1}). \tag{3}$$

The quantity $\mathbb{E}[\log p(Y_{>i} \mid C_i)]$ is estimated using Eq. 4, with Y fixed to the final answer generated in the initial trajectory. Here, $\mathcal{L}(C_i, Y_{>i})$ denotes the sequence-level loss of the complete sequence $C_i\|Y_{>i}$, while $\log p(C_i) = -\mathcal{L}(C_i)$ denotes the log-probability of the reasoning prefix. Rollout trajectories that do not produce the same final answer Y are excluded from the estimation. The detailed derivation is provided in Appendix A.2.

$$\mathbb{E}\left[\log \widehat{p(Y_{>i}} \mid C_i)\right] = -\frac{1}{N_{\text{succ}}} \sum_{n=1}^{N} \mathbf{1}\left[Y^{(n)} = Y\right] \mathcal{L}\left(C_i, Y_{>i}^{(n)}\right) + \mathcal{L}(C_i), \tag{4}$$

where

$$N_{\text{succ}} = \sum_{n=1}^{N} \mathbf{1}\left[Y^{(n)} = Y\right]. \tag{5}$$

3.3 Confidence Estimation via Evidence Information Gain

CoT reasoning can improve LLM accuracy Wei et al. (2022) and provide signals for confidence estimation. Since evidence-level information gain captures each evidence segment's contribution toward the final answer, its sequence provides a measure of reasoning informativeness. We therefore use these information gains as a **novel signal** for estimating confidence in the final answer.

Information-gain trajectories may exhibit patterns associated with reasoning uncertainty, such as late surges or substantial fluctuations across evidence segments. In contrast, more consistent trajectories may indicate greater confidence. However, these patterns should not be assumed to transfer directly from human decision-making to LLM reasoning, whose evidence-accumulation dynamics may differ substantially.

In the previous section 3.2, we introduced a method for estimating the information gain of each evidence, where $\mathbb{E}[\log p(Y_{>i} \mid C_i)]$ is estimated using the LLM's loss on the generated response. Alternatively, we estimate $p\left(Y_{>i} \mid C_i\right)$ via self-consistency by sampling multiple continuations and use the resulting probability as a single estimate of $\mathbb{E}\left[\log p(Y_{>i} \mid C_i)\right]$. While the loss-based approach may provide a more accurate estimate, self-consistency is a **black-box** method that does not require access to the model's loss or token-level probabilities. In addition to loss- and self-consistency-based estimates, we use their changes between consecutive evidence segments as additional features. For each evidence i, we construct the feature tuple defined in Eq. 6:

$$\mathbf{f}_i = \left(e^i_{\text{loss}},\, e^i_{\text{sc}},\, \Delta^i_{\text{loss}},\, \Delta^i_{\text{sc}}\right), \qquad \Delta^i_{\text{loss}} = e^i_{\text{loss}} - e^{i-1}_{\text{loss}}, \qquad \Delta^i_{\text{sc}} = e^i_{\text{sc}} - e^{i-1}_{\text{sc}}. \tag{6}$$

$$\mathbf{F} = [\mathbf{f}_1; \mathbf{f}_2; \ldots; \mathbf{f}_k] \in \mathbb{R}^{4k}. \tag{7}$$

The CoT trace is partitioned into equal-length evidence segments, from which four predefined features are extracted and concatenated into $\mathbf{F}$. Logistic regression then maps $\mathbf{F}$ to the probability of a correct final answer, which serves as the confidence estimate.

4 Experiments

4.1 Experimental Setup

Models and Benchmarks. We evaluate our confidence estimation method on two open-weight reasoning models, DeepSeek-R1-Distill-Qwen-7B Guo et al. (2025) and Qwen3-8B Yang et al. (2025), across seven benchmarks spanning mathematical, scientific, factual, and general reasoning. The suite includes AIME, MATH500, and GSM8K for mathematical reasoning; GPQA, MMLU, and MMLU-Pro for scientific and academic knowledge; TruthfulQA for factuality.

Evaluation Metrics. We evaluate confidence quality using **AUROC**, which measures discrimination between correct and incorrect predictions, and **Expected Calibration Error (ECE)**, which measures the alignment between predicted confidence and empirical accuracy Guo et al. (2017); Kadavath et al. (2022). Non-probabilistic confidence scores are min–max normalized before computing ECE.

4.2 Baselines and Proposed Method

Token-level baselines. The token-level baselines compute statistics over the CoT reasoning trace. Specifically, the mean and sum of token-level loss (**Mean-Loss**, **Sum-Loss**), entropy (**Mean-Ent**, **Sum-Ent**), and token probability (**Mean-Prob**, **Sum-Prob**) are considered.

Representation-based baseline. **LastRep** uses the final-layer representation of the CoT as input to a logistic regression classifier trained to predict answer correctness. The classifier's output probability is used as the confidence estimate.

Self-consistency baselines. **SC-10** generates 10 independent rollouts for each question, with confidence estimated from the agreement among their final answers. To provide a computationally matched comparison, **SC-Budget** increases the self-consistency sampling budget to 25, 50, 75, 100, and 125 rollouts for $K \in \{5, 10, 15, 20, 25\}$, respectively, where K denotes the number of evidences in the chain-of-thought reasoning trace. These budgets approximately match the computational cost of our method with 10 rollouts per evidence.

Proposed method. The proposed method, denoted as **CoT-EIG**, first generates the complete response with a temperature of 0.2 and divides the CoT into K consecutive evidences. For each evidence, 10 rollout continuations are generated using a temperature of 1.0 to extract information-gain-based features. Two variants based on token aggregation are considered: **CoT-EIG-Mean**, which averages the features over tokens, and **CoT-EIG-Sum**, which sums them.

Table 1: Confidence estimation performance of different baselines and CoT-EIG on Qwen3-8B and DeepSeek-R1-Distill-Qwen-7B using 5 CoT evidences. Results are mean ± standard deviation over 50 bootstrap resamples of the held-out evaluation set, drawn with replacement.

Method	Qwen3-8B		DeepSeek-R1-Distill-Qwen-7B	
	AUROC ↑	ECE ↓	AUROC ↑	ECE ↓
Token-level baselines				
Mean-Ent	0.61 ± 0.02	0.22 ± 0.02	0.58 ± 0.02	0.17 ± 0.02
Sum-Ent	0.62 ± 0.02	0.22 ± 0.01	0.60 ± 0.02	0.24 ± 0.01
Mean-Loss	0.61 ± 0.02	0.23 ± 0.02	0.57 ± 0.02	0.16 ± 0.02
Sum-Loss	0.62 ± 0.02	0.22 ± 0.01	0.60 ± 0.02	0.24 ± 0.01
Mean-Prob	0.58 ± 0.02	0.17 ± 0.01	0.58 ± 0.02	0.23 ± 0.01
Representation-based baseline				
LastRep	0.60 ± 0.02	0.28 ± 0.01	0.58 ± 0.02	0.31 ± 0.01
Self-consistency baselines				
SC-10	0.69 ± 0.02	0.18 ± 0.02	0.71 ± 0.02	0.16 ± 0.01
SC-Budget	0.78 ± 0.02	0.15 ± 0.01	—	—
Proposed method				
CoT-EIG-Mean	$\mathbf{0.84 \pm 0.01}$	$\mathbf{0.11 \pm 0.01}$	0.83 ± 0.01	$\mathbf{0.08 \pm 0.01}$
CoT-EIG-Sum	0.81 ± 0.02	0.12 ± 0.01	$\mathbf{0.84 \pm 0.01}$	0.13 ± 0.01

4.3 Cross-Benchmark Generalization

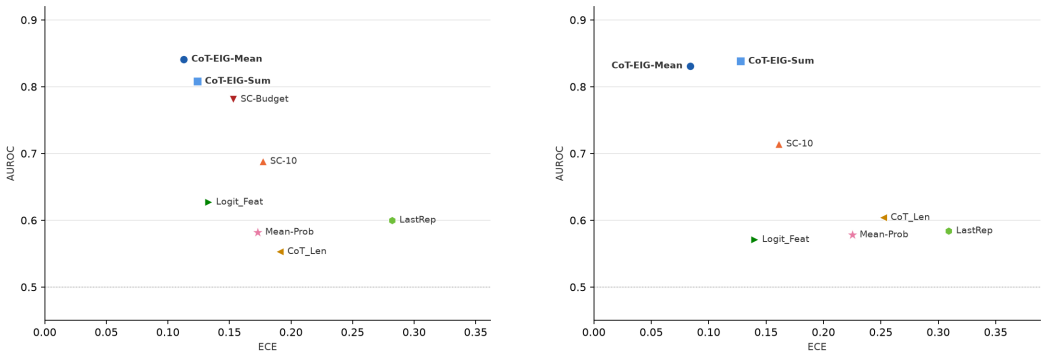
We evaluate the generalization of confidence estimation using a domain-based holdout protocol. For each target benchmark, the estimator is trained on the remaining benchmarks and evaluated on the held-out benchmark. For example, MATH500, GSM8K, and AIME form the mathematical reasoning domain, while MMLU and MMLU-Pro form the multiple-choice knowledge domain. Each benchmark is used once as the target, and results are averaged across all holdout evaluations.All experiments are repeated 50 times using bootstrap resampling of the held-out evaluation set with replacement. Given the low sampling temperature 0.2 of the initial response, repeated model generations are not used as an additional source of variability; instead, uncertainty is estimated through bootstrap resampling. The mean and standard deviation are reported across the 50 resamples.

Overall, the CoT-EIG variants consistently achieve higher AUROC and lower ECE than the token-level and self-consistency baselines, as shown in Figure 3. On Qwen3-8B, CoT-EIG-Mean and CoT-EIG-Sum achieve AUROC values of 0.84 and 0.81, with ECE values of 0.11 and 0.12, respectively. On DeepSeek-R1-Distill-Qwen-7B, they achieve AUROC values of 0.83 and 0.84, with ECE values of 0.08 and 0.13, respectively. Detailed results for individual holdout evaluations are provided in the Appendix B.

4.4 Analysis of Feature Contributions

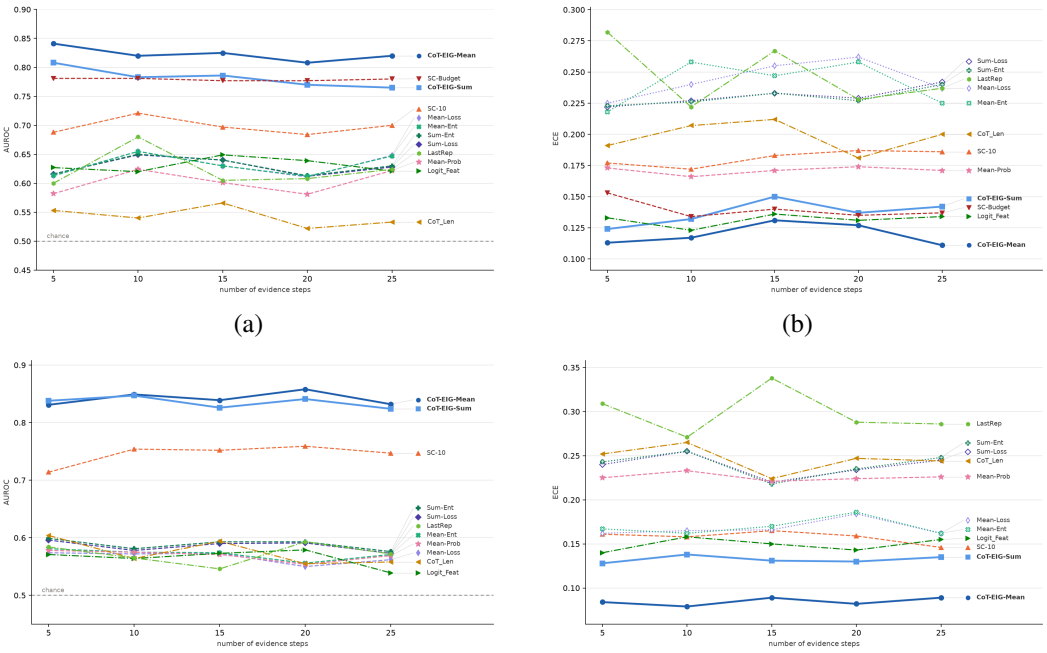
To better understand how the proposed features contribute to confidence estimation, the estimated confidence scores on the test set are divided into two classes: *high confidence* for probabilities ≥ 0.5 and *low confidence* for probabilities < 0.5. A decision tree is then trained on the proposed features to predict these confidence labels, providing an interpretable view of how individual features contribute to the estimated confidence. The tree is visualized up to depth four. For this analysis, **CoT-EIG-Mean** with $K = 5$ evidences is considered.

As shown in Figure 2, features from multiple evidences are used as splitting criteria, indicating that confidence estimation relies on information distributed across the reasoning trace. Notably, as shown in Table 2, self-consistency-based features have the highest aggregate feature importance across the internal split nodes, compared with loss-based features, indicating that they contribute more strongly to distinguishing between high- and low-confidence predictions. The non-zero importance of loss-based features indicates that they also contribute to confidence estimation.



(a) Qwen3-8B (b) DeepSeek-R1-Distill-Qwen-7B

Figure 3: Confidence estimation performance of the baselines and proposed methods in terms of AUROC and ECE. Each point represents one method, with results shown separately for (a) Qwen3-8B and (b) DeepSeek-R1-Distill-Qwen-7B. Methods located higher and farther to the left correspond to higher AUROC and lower ECE, respectively, and thus indicate better confidence estimation performance.



(a) (b)

(c) (d)

Figure 4: Performance of different confidence estimation baselines across varying numbers of evidences. The top row reports results for Qwen3-8B, while the bottom row reports results for DeepSeek-R1-Distill-Qwen-7B. The left and right columns show AUROC and ECE, respectively.

5 Ablation Study

5.1 Effect of the Number of CoT Evidences

We first investigate how the number of CoT evidences affects the performance of our confidence estimation method. Specifically, we divide each reasoning trace into 5, 10, 15, 20, or 25 evidences and recompute all features used by our confidence estimator for each setting. This analysis allows us to assess whether the proposed approach is sensitive to the granularity at which the reasoning process

is decomposed. By varying the number of evidences while keeping the remaining components of the method unchanged, we characterize the effect of evidence granularity on confidence estimation performance.

Figure 4 shows the effect of varying the number of CoT evidences on confidence estimation performance. Although both AUROC and ECE exhibit minor fluctuations across different numbers of evidences, the overall results remain relatively stable. This suggests that increasing or decreasing the number of evidences does not substantially affect the confidence estimation performance across the evaluated settings. Since the evaluation covers a diverse set of benchmarks, detailed results for each individual benchmark are provided in Appendix B.

5.2 Effect of Computational Budget

We next examine whether the performance gains persist under a comparable computational budget. To this end, our method is compared with the computationally matched self-consistency baseline, **SC-Budget**, whose rollout budget is adjusted according to the number of evidences, denoted by K, to approximately match the computational cost of our method. Since self-consistency generally benefits from additional samples, this comparison evaluates whether the gains of our approach persist when the baseline is given a substantially larger sampling budget.

Figure 4 shows that, on Qwen3-8B, **SC-Budget** performs worse than **CoT-EIG-Mean** and achieves performance comparable to **CoT-EIG-Sum**. These results suggest that **CoT-EIG-Mean** provides a meaningful improvement over self-consistency even under a comparable computational budget.

5.3 Ablation Study of Feature Sets

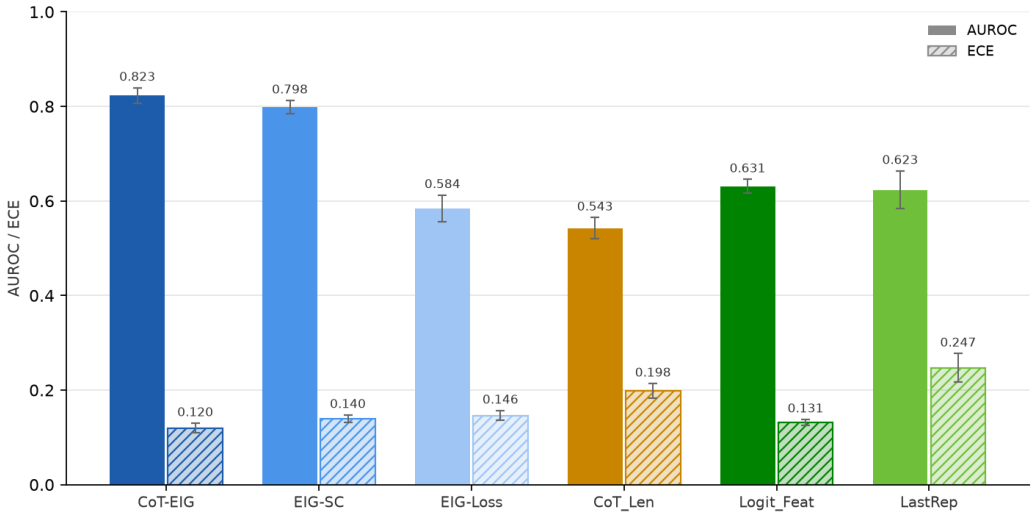
An ablation study is conducted to examine the contribution of different feature sets to confidence estimation.Six feature sets are considered: **CoT-EIG**, which uses all four proposed features; **EIG-SC**, which uses only the self-consistency-based features; **EIG-Loss**, which uses only the loss-based features; **Logit_Feat**, which consists of conventional token-level features, including entropy, loss, and probability statistics; **LastRep**, which uses the final-layer representation of the reasoning trace; and **CoT_Len**, which uses the length of the generated response as the feature. All feature sets are used to train the same logistic regression confidence estimator, allowing the effect of the underlying feature representation to be isolated.

Although **CoT-EIG** already combines both loss- and self-consistency (SC)-based features, it achieves the strongest overall performance, with improvements in AUROC and reductions in ECE, the EIG-SC variant performs very close to this full feature set, as shown in Figure 5. In contrast, EIG-Loss does not exhibit a comparable performance. Within the CoT-EIG framework, incorporating loss features provides a modest improvement in ECE while either improving or maintaining AUROC across the evaluated settings. These results suggest that SC-based features can provide performance close to the full CoT-EIG feature set when the underlying LLM is accessible only as a black box.

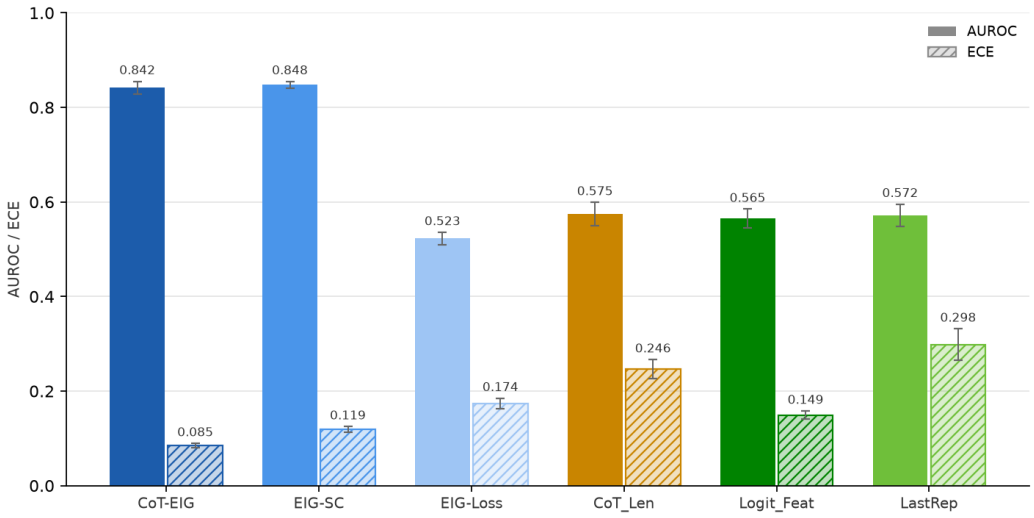
5.4 Conclusion

In this work, we estimate LLM confidence by modeling Chain-of-Thought as sequential evidence toward the final answer. Inspired by the Diffusion Decision Model, we quantify the information gain of reasoning steps and use the resulting evidence-accumulation patterns as confidence features. Across seven benchmarks and two reasoning models, our approach consistently improves AUROC and reduces ECE compared with existing methods, while maintaining comparable performance in both white-box and black-box settings, showing that reasoning trajectories provide effective confidence signals without requiring access to model internals.

More broadly, confidence estimation is an important metacognitive capability of LLMs and should ideally be assessed independently of their task-solving ability. By modeling confidence continuously throughout the reasoning trajectory, trajectory-level confidence estimation may capture evolving uncertainty that is missed by a single final confidence score. This perspective motivates further research into the temporal dynamics of confidence and its relationship to prediction correctness.



(a) Qwen3-8B



(b) DeepSeek-R1-Distill-Qwen-7B

Figure 5: Ablation study of different feature sets for confidence estimation on two language models. The top and bottom panels report results for Qwen3-8B and DeepSeek-R1-Distill-Qwen-7B, respectively.

AI USE STATEMENT

We used generative AI tools primarily as an alternative to conventional internet search. Information obtained through these tools was subsequently reviewed and independently verified against relevant sources to validate its accuracy.

The authors take full responsibility for the final content of this work, including all text, claims, analyses, and research artifacts, regardless of whether generative AI tools were used in their preparation.

ETHICS STATEMENT

This work uses publicly available datasets and pretrained language models and does not involve human subjects, personally identifiable information, or sensitive data. We are not aware of significant

ethical concerns arising from this work. By estimating confidence in LLM outputs, our approach may help reduce concerns associated with over-reliance on potentially incorrect or uncertain outputs by making uncertainty more explicit.

Reproducibility statement

To support reproducibility, we conducted the experiments with multiple independent runs and evaluated the proposed method across multiple models and benchmarks. The main hyperparameters and experimental settings are reported in the paper, with additional implementation details provided in the supplementary materials. The complete source code, experimental outputs, and information required to access the datasets are provided in the accompanying GitHub repository.

References

Zijun Chen, Wenbo Hu, and Richang Hong. Deep hidden cognition facilitates reliable chain-of-thought reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 37295–37304, 2026. doi: 10.1609/aaai.v40i44.41061.

Siddhartha Devic, Charlotte Peale, Arwen Bradley, Sinead Williamson, Preetum Nakkiran, and Aravind Gollakota. Trace length is a simple uncertainty signal in reasoning models. *arXiv preprint arXiv:2510.10409*, 2025.

Yichao Fu, Xuewei Wang, Hao Zhang, Yuandong Tian, and Jiawei Zhao. Deep think with confidence. In *International Conference on Learning Representations*, 2026.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 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)*, pp. 6577–6595. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.366.

Nils Grünefeld, Bertram Højer, Philipp Mondorf, Barbara Plank, Anna Rogers, Christian Hardmeier, Stefan Heinrich, and Jes Frellsen. Tracing uncertainty in language model "reasoning". *arXiv preprint arXiv:2605.07776*, 2026. doi: 10.48550/arXiv.2605.07776.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, 2017.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Hasan Abed Al Kader Hammoud, Hani Itani, and Bernard Ghanem. Beyond the last answer: Your reasoning trace uncovers more than you think. *arXiv preprint arXiv:2504.20708*, 2025.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 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, 2021. doi: 10.1162/tacl_a_00407.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Samuel Bowman, Stanislav Fort, Deep Ganguli, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Zhewei Kang, Xuandong Zhao, and Dawn Song. Scalable best-of-n selection for large language models via self-certainty. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*, 2023.

Xiaoou Liu, Tiejin Chen, Dengjia Zhang, Yaqing Wang, Lu Cheng, and Hua Wei. Diagnosing multi-step reasoning failures in black-box llms via stepwise confidence attribution. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Zhenjiang Mao, Artem Bisliouk, Rohith Reddy Nama, and Ivan Ruchkin. Temporalizing confidence: Evaluation of chain-of-thought reasoning with signal temporal logic. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pp. 882–890, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.bea-1.65.

Zhenjiang Mao, Anirudhh Venkat, Artem Bisliouk, Sindhura Kumbakonam Subramanian, Akshat Kothiyal, Saithej Singhu, and Ivan Ruchkin. Confidence over time: Confidence calibration with temporal logic for large language model reasoning. In *Findings of the Association for Computational Linguistics: ACL 2026*, 2026.

Jakub Podolak and Rajeev Verma. Read your own mind: Reasoning helps surface self-confidence signals in llms. In *Proceedings of the 2nd Workshop on Uncertainty-Aware NLP*, pp. 247–258. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.uncertainlp-main.21.

Roger Ratcliff. A theory of memory retrieval. *Psychological Review*, 85(2):59–108, 1978. doi: 10.1037/0033-295X.85.2.59.

Roger Ratcliff and Gail McKoon. The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, 20(4):873–922, 2008. doi: 10.1162/neco.2008.12-06-420.

Roger Ratcliff and Francis Tuerlinckx. Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review*, 9(3):438–481, 2002. doi: 10.3758/BF03196302.

Roger Ratcliff, Philip L. Smith, Scott D. Brown, and Gail McKoon. Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences*, 20(4):260–281, 2016. doi: 10.1016/j.tics.2016.01.007.

Ali Razghandi, Seyed Mohammad Hadi Hosseini, and Mahdieh Soleymani Baghshah. CER: Confidence enhanced reasoning in LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7918–7938, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.390.

Dennis Ulmer, Martin Gubri, Hwaran Lee, Sangdoo Yun, and Seong Joon Oh. Calibrating large language models using their generations only. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15440–15459. Association for Computational Linguistics, 2024.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837. Curran Associates, Inc., 2022. doi: 10.52202/068431-1800.

Juming Xiong, Kevin Guo, Congning Ni, Weixin Liu, Chao Yan, et al. Learning when to sample: Confidence-aware selective sampling for efficient chain-of-thought reasoning. *arXiv preprint arXiv:2603.08999*, 2026.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *International Conference on Learning Representations*, 2024.

An Yang, An Yang, Baosong Zhang, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Table 2: Feature importance based on the fitted decision tree, with higher-importance evidence–feature pairs reported in descending order. The decision tree is fitted on the combined data from all benchmarks evaluated using Qwen3-8B.

Evidence	Feature	Importance
1	self-consistency	0.6393
3	self-consistency	0.1184
3	loss	0.0847
4	self-consistency	0.0539
1	loss	0.0304
2	delta_loss	0.0219
5	loss	0.0094
2	delta_self-consistency	0.0084
4	delta_self-consistency	0.0083
3	delta_loss	0.0078
4	delta_loss	0.0070
2	loss	0.0056
5	self-consistency	0.0028
5	delta_self-consistency	0.0020

A PROOFS

A.1

$$\begin{aligned}\mathbb{E}[\log p(Y \mid C_i)] - \mathbb{E}[\log p(Y \mid C_{i-1})] &= \mathbb{E}\left[\log \frac{p(Y \mid C_i)}{p(Y \mid C_{i-1})}\right] \\ &= \mathbb{E}\left[\log \frac{p(Y \mid C_i, C_{i-1})}{p(Y \mid C_{i-1})}\right] \\ &= \mathbb{E}\left[\log \frac{p(Y, C_i \mid C_{i-1})}{p(Y \mid C_{i-1})\, p(C_i \mid C_{i-1})}\right] \\ &= I(Y; C_i \mid C_{i-1})\,.\end{aligned}$$

A.2

By the chain rule:

$$\mathbb{E}\left[\log p(Y_{>i} \mid C_i)\right] = \mathbb{E}\left[\log \frac{p(C_i, Y_{>i})}{p(C_i)}\right].$$

Thus,

$$\begin{aligned}\mathbb{E}\left[\log p(Y_{>i} \mid C_i)\right] &= \mathbb{E}\left[\log p(C_i, Y_{>i}) - \log p(C_i)\right] \\ &= \mathbb{E}\left[\log p(C_i, Y_{>i})\right] - \log p(C_i).\end{aligned}$$

Using the sequence-level losses

$$\mathcal{L}(C_i, Y_{>i}) = -\log p(C_i, Y_{>i}), \qquad \mathcal{L}(C_i) = -\log p(C_i),$$

we obtain

$$\mathbb{E}\left[\log p(Y_{>i} \mid C_i)\right] = -\mathbb{E}\left[\mathcal{L}(C_i, Y_{>i})\right] + \mathcal{L}(C_i).$$

Finally, using N independently sampled continuations $\{Y_{>i}^{(n)}\}_{n=1}^{N}$, we obtain

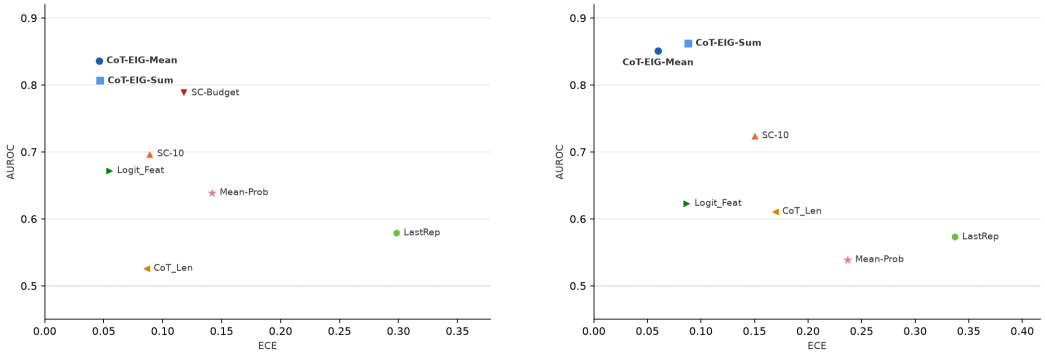
$$\mathbb{E}\left[\log \widehat{p(Y_{>i}} \mid C_i)\right] = -\frac{1}{N}\sum_{n=1}^{N} \mathcal{L}\left(C_i, Y_{>i}^{(n)}\right) + \mathcal{L}(C_i).$$

B Results for Each Held-Out Group

B.1 Multiple-Choice Knowledge Domain

Table 3: Same as Table 1, with the multiple-choice knowledge domain (MMLU and MMLU-Pro) held out.

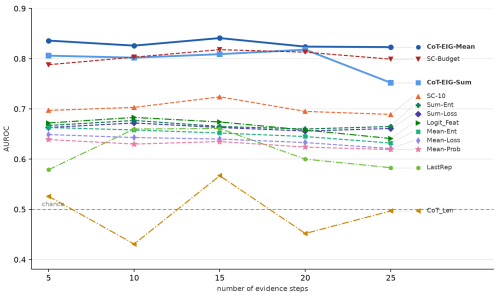
Method	Qwen3-8B AUROC ↑	Qwen3-8B ECE ↓	DeepSeek-R1-Distill-Qwen-7B AUROC ↑	DeepSeek-R1-Distill-Qwen-7B ECE ↓
Token-level baselines				
Mean-Ent	0.66 ± 0.02	0.20 ± 0.02	0.54 ± 0.03	0.14 ± 0.03
Sum-Ent	0.67 ± 0.02	0.21 ± 0.02	0.59 ± 0.02	0.27 ± 0.02
Mean-Loss	0.65 ± 0.02	0.20 ± 0.04	0.54 ± 0.03	0.12 ± 0.04
Sum-Loss	0.66 ± 0.02	0.21 ± 0.02	0.58 ± 0.02	0.28 ± 0.02
Mean-Prob	0.64 ± 0.03	0.14 ± 0.02	0.54 ± 0.03	0.24 ± 0.02
Representation-based baseline				
LastRep	0.58 ± 0.02	0.30 ± 0.02	0.57 ± 0.02	0.34 ± 0.02
Self-consistency baselines				
SC-10	0.70 ± 0.02	0.09 ± 0.02	0.72 ± 0.02	0.15 ± 0.02
SC-Budget	0.79 ± 0.02	0.12 ± 0.02	—	—
Proposed method				
CoT-EIG-Mean	$\mathbf{0.84 \pm 0.02}$	$\mathbf{0.05 \pm 0.01}$	0.85 ± 0.02	$\mathbf{0.06 \pm 0.01}$
CoT-EIG-Sum	0.81 ± 0.02	$\mathbf{0.05 \pm 0.01}$	$\mathbf{0.86 \pm 0.01}$	0.09 ± 0.01



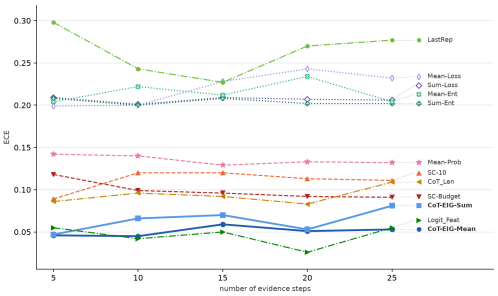
(a) Qwen3-8B

(b) DeepSeek-R1-Distill-Qwen-7B

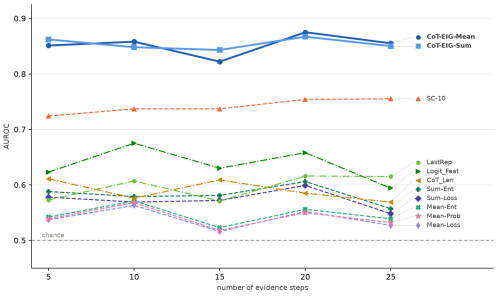
Figure 6: Same as Figure 3, with the multiple-choice knowledge domain (MMLU and MMLU-Pro) held out.



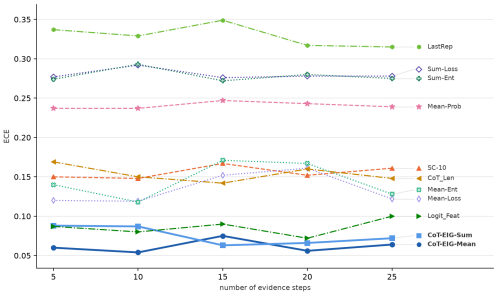
(a) Qwen3-8B, AUROC



(b) Qwen3-8B, ECE

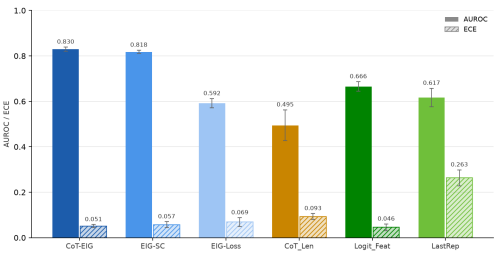


(c) DeepSeek-R1-Distill-Qwen-7B, AUROC

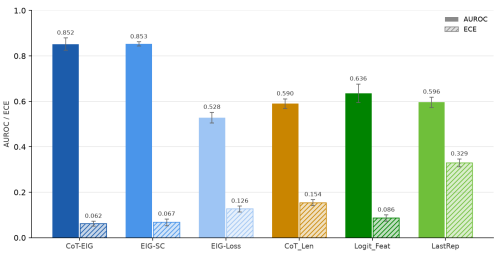


(d) DeepSeek-R1-Distill-Qwen-7B, ECE

Figure 7: Same as Figure 4, with the multiple-choice knowledge domain (MMLU and MMLU-Pro) held out.



(a) Qwen3-8B



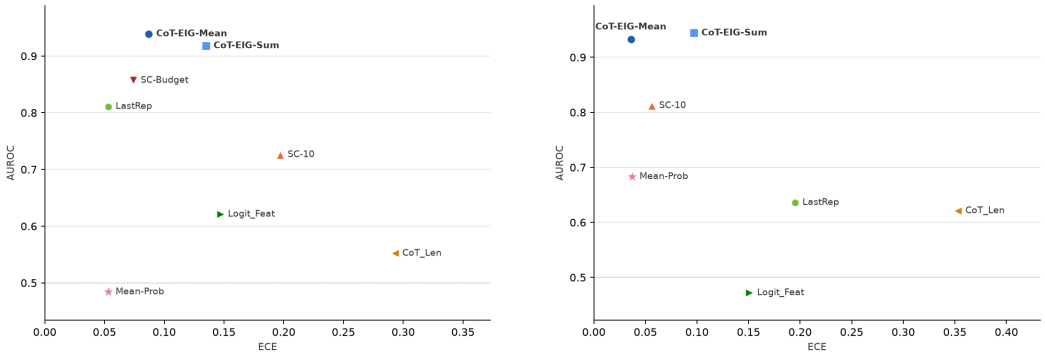
(b) DeepSeek-R1-Distill-Qwen-7B

Figure 8: Same as Figure 5, with the multiple-choice knowledge domain (MMLU and MMLU-Pro) held out.

B.2 MATHEMATICS DOMAIN

Table 4: Same as Table 1, with the mathematics domain (GSM8K, MATH500 and AIME) held out.

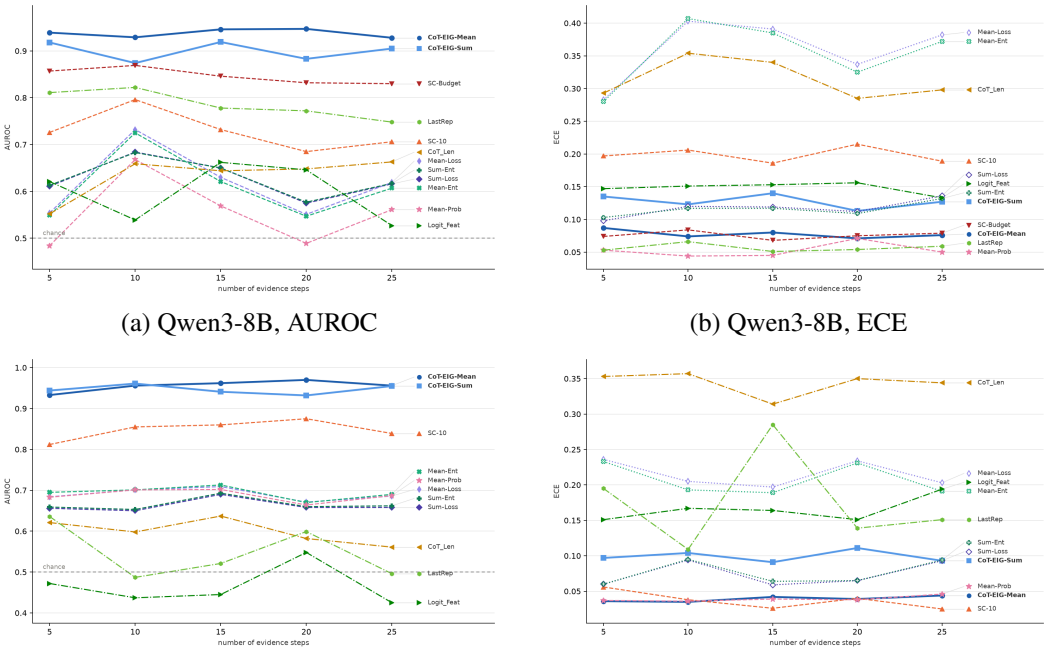
Method	Qwen3-8B AUROC ↑	Qwen3-8B ECE ↓	DeepSeek-R1-Distill-Qwen-7B AUROC ↑	DeepSeek-R1-Distill-Qwen-7B ECE ↓
Token-level baselines				
Mean-Ent	0.55 ± 0.06	0.28 ± 0.06	0.69 ± 0.05	0.23 ± 0.02
Sum-Ent	0.61 ± 0.04	0.10 ± 0.02	0.66 ± 0.05	0.06 ± 0.01
Mean-Loss	0.55 ± 0.06	0.28 ± 0.06	0.68 ± 0.05	0.24 ± 0.02
Sum-Loss	0.61 ± 0.05	0.10 ± 0.02	0.66 ± 0.05	0.06 ± 0.01
Mean-Prob	0.48 ± 0.06	$\mathbf{0.05 \pm 0.01}$	0.68 ± 0.05	$\mathbf{0.04 \pm 0.01}$
Representation-based baseline				
LastRep	0.81 ± 0.04	$\mathbf{0.05 \pm 0.01}$	0.64 ± 0.04	0.20 ± 0.02
Self-consistency baselines				
SC-10	0.73 ± 0.04	0.20 ± 0.04	0.81 ± 0.04	0.06 ± 0.02
SC-Budget	0.86 ± 0.03	0.07 ± 0.02	—	—
Proposed method				
CoT-EIG-Mean	$\mathbf{0.94 \pm 0.03}$	0.09 ± 0.01	0.93 ± 0.03	$\mathbf{0.04 \pm 0.01}$
CoT-EIG-Sum	0.92 ± 0.03	0.14 ± 0.01	$\mathbf{0.94 \pm 0.02}$	0.10 ± 0.01



(a) Qwen3-8B

(b) DeepSeek-R1-Distill-Qwen-7B

Figure 9: Same as Figure 3, with the mathematics domain (GSM8K, MATH500 and AIME) held out.



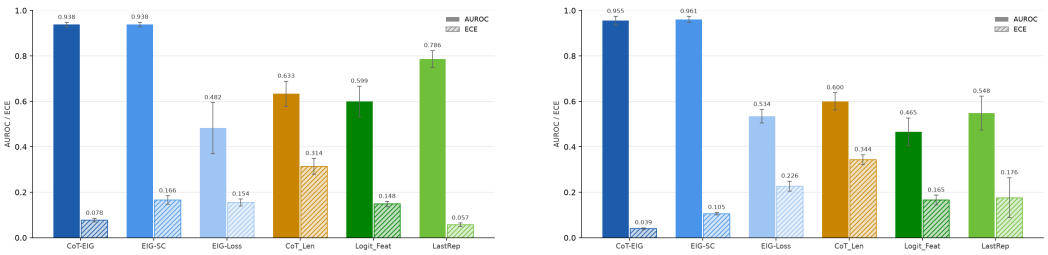
(a) Qwen3-8B, AUROC

(b) Qwen3-8B, ECE

(c) DeepSeek-R1-Distill-Qwen-7B, AUROC

(d) DeepSeek-R1-Distill-Qwen-7B, ECE

Figure 10: Same as Figure 4, with the mathematics domain (GSM8K, MATH500 and AIME) held out.



(a) Qwen3-8B

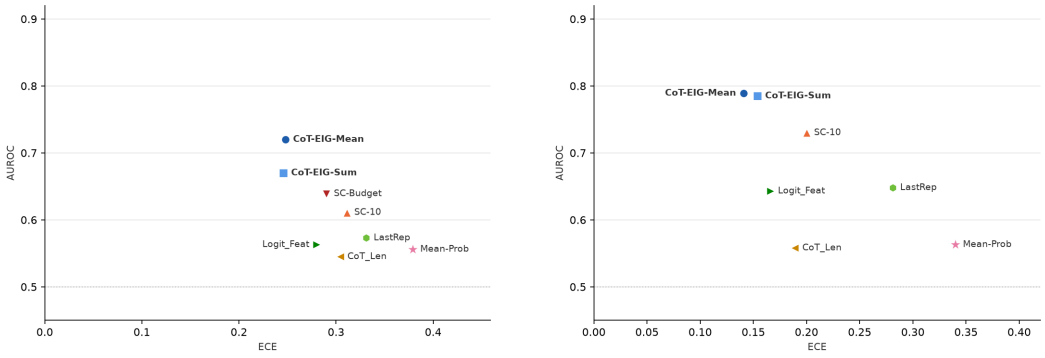
(b) DeepSeek-R1-Distill-Qwen-7B

Figure 11: Same as Figure 5, with the mathematics domain (GSM8K, MATH500 and AIME) held out.

B.3 GPQA

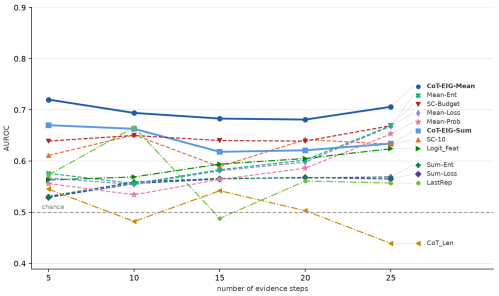
Table 5: Same as Table 1, with GPQA held out.

Method	Qwen3-8B AUROC ↑	Qwen3-8B ECE ↓	DeepSeek-R1-Distill-Qwen-7B AUROC ↑	DeepSeek-R1-Distill-Qwen-7B ECE ↓
Token-level baselines				
Mean-Ent	0.58 ± 0.04	$\mathbf{0.14 \pm 0.03}$	0.56 ± 0.05	0.17 ± 0.03
Sum-Ent	0.53 ± 0.04	0.43 ± 0.04	0.60 ± 0.05	0.23 ± 0.03
Mean-Loss	0.57 ± 0.04	0.16 ± 0.03	0.55 ± 0.05	0.17 ± 0.04
Sum-Loss	0.53 ± 0.04	0.43 ± 0.04	0.60 ± 0.05	0.22 ± 0.03
Mean-Prob	0.56 ± 0.04	0.38 ± 0.04	0.56 ± 0.05	0.34 ± 0.04
Representation-based baseline				
LastRep	0.57 ± 0.04	0.33 ± 0.03	0.65 ± 0.04	0.28 ± 0.04
Self-consistency baselines				
SC-10	0.61 ± 0.04	0.31 ± 0.05	0.73 ± 0.04	0.20 ± 0.03
SC-Budget	0.64 ± 0.04	0.29 ± 0.04	—	—
Proposed method				
CoT-EIG-Mean	$\mathbf{0.72 \pm 0.04}$	0.25 ± 0.03	$\mathbf{0.79 \pm 0.03}$	$\mathbf{0.14 \pm 0.03}$
CoT-EIG-Sum	0.67 ± 0.04	0.25 ± 0.04	$\mathbf{0.79 \pm 0.03}$	0.15 ± 0.03

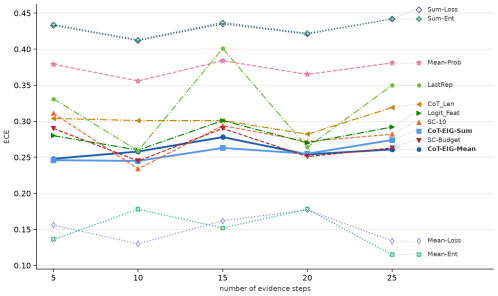


(a) Qwen3-8B (b) DeepSeek-R1-Distill-Qwen-7B

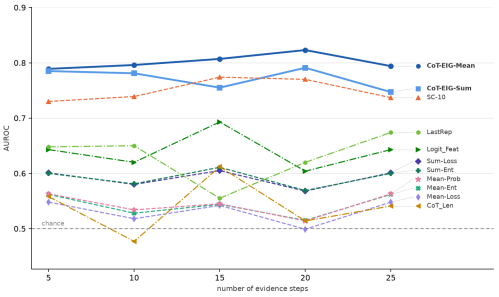
Figure 12: Same as Figure 3, with GPQA held out.



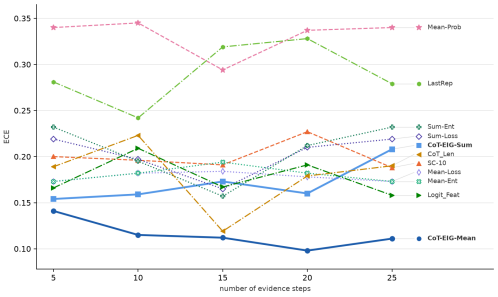
(a) Qwen3-8B, AUROC



(b) Qwen3-8B, ECE

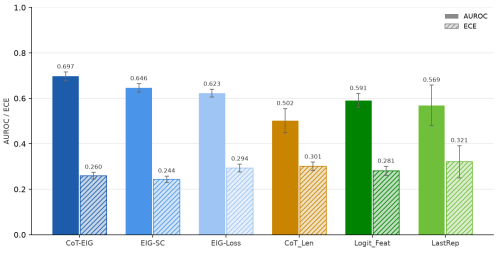


(c) DeepSeek-R1-Distill-Qwen-7B, AUROC

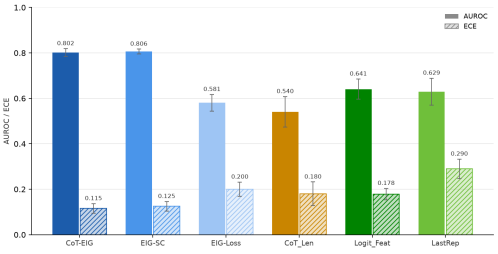


(d) DeepSeek-R1-Distill-Qwen-7B, ECE

Figure 13: Same as Figure 4, with GPQA held out.



(a) Qwen3-8B



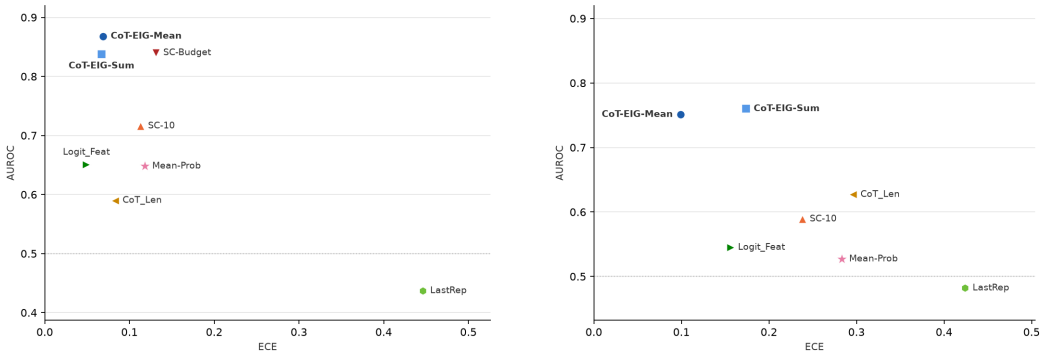
(b) DeepSeek-R1-Distill-Qwen-7B

Figure 14: Same as Figure 5, with GPQA held out.

B.4 TRUTHFULQA

Table 6: Same as Table 1, with TruthfulQA held out.

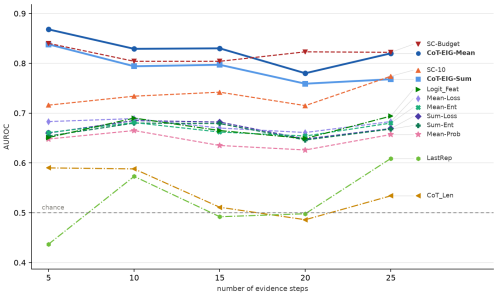
Method	Qwen3-8B AUROC ↑	Qwen3-8B ECE ↓	DeepSeek-R1-Distill-Qwen-7B AUROC ↑	DeepSeek-R1-Distill-Qwen-7B ECE ↓
Token-level baselines				
Mean-Ent	0.66 ± 0.03	0.25 ± 0.03	0.53 ± 0.03	0.12 ± 0.03
Sum-Ent	0.65 ± 0.03	0.15 ± 0.02	0.55 ± 0.03	0.41 ± 0.03
Mean-Loss	0.68 ± 0.03	0.26 ± 0.02	0.53 ± 0.03	0.12 ± 0.03
Sum-Loss	0.66 ± 0.03	0.15 ± 0.02	0.55 ± 0.03	0.41 ± 0.03
Mean-Prob	0.65 ± 0.03	0.12 ± 0.02	0.53 ± 0.03	0.28 ± 0.03
Representation-based baseline				
LastRep	0.44 ± 0.04	0.45 ± 0.03	0.48 ± 0.03	0.42 ± 0.03
Self-consistency baselines				
SC-10	0.72 ± 0.03	0.11 ± 0.02	0.59 ± 0.03	0.24 ± 0.03
SC-Budget	0.84 ± 0.02	0.13 ± 0.02	—	—
Proposed method				
CoT-EIG-Mean	$\mathbf{0.87 \pm 0.02}$	$\mathbf{0.07 \pm 0.02}$	0.75 ± 0.03	$\mathbf{0.10 \pm 0.02}$
CoT-EIG-Sum	0.84 ± 0.02	$\mathbf{0.07 \pm 0.02}$	$\mathbf{0.76 \pm 0.03}$	0.17 ± 0.02



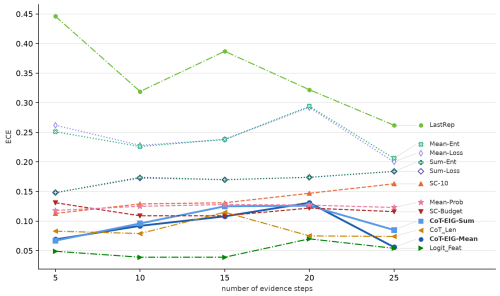
(a) Qwen3-8B

(b) DeepSeek-R1-Distill-Qwen-7B

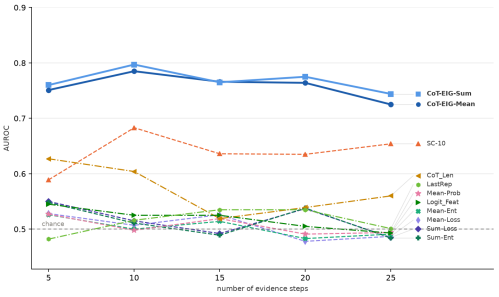
Figure 15: Same as Figure 3, with TruthfulQA held out.



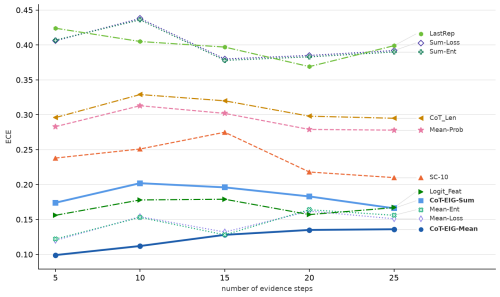
(a) Qwen3-8B, AUROC



(b) Qwen3-8B, ECE

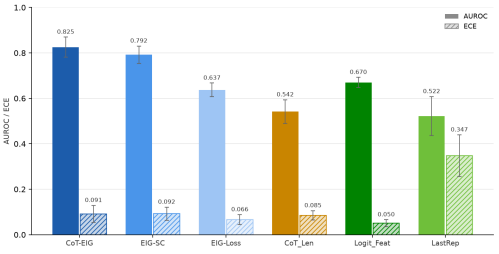


(c) DeepSeek-R1-Distill-Qwen-7B, AUROC

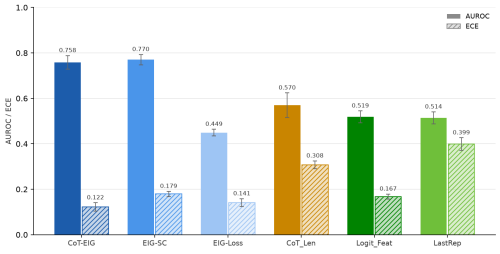


(d) DeepSeek-R1-Distill-Qwen-7B, ECE

Figure 16: Same as Figure 4, with TruthfulQA held out.



(a) Qwen3-8B



(b) DeepSeek-R1-Distill-Qwen-7B

Figure 17: Same as Figure 5, with TruthfulQA held out.