Chain-of-Thought Prompting Obscures Hallucination Cues in Large Language Models: An Empirical Evaluation

Jiahao Cheng[1] **Tiancheng Su**[1] **Jia Yuan**[1] **Guoxiu He**[1]*
Jiawei Liu[2] **Xinqi Tao**[3] **Jingwen Xie**[3] **Huaxia Li**[3]

[1] East China Normal University, Shanghai, China
[2]Wuhan University, Wuhan, China
[3]Xiaohongshu Inc., Shanghai, China
chengjiahao@stu.ecnu.edu.cn, gxhe@fem.ecnu.edu.cn

Abstract

Large Language Models (LLMs) often exhibit *hallucinations*, generating factually incorrect or semantically irrelevant content in response to prompts. Chain-of-Thought (CoT) prompting can mitigate hallucinations by encouraging step-by-step reasoning, but its impact on hallucination detection remains underexplored. To bridge this gap, we conduct a systematic empirical evaluation. We begin with a pilot experiment, revealing that CoT reasoning significantly affects the LLM's internal states and token probability distributions. Building on this, we evaluate the impact of various CoT prompting methods on mainstream hallucination detection methods across both instruction-tuned and reasoning-oriented LLMs. Specifically, we examine three key dimensions: changes in hallucination score distributions, variations in detection accuracy, and shifts in detection confidence. Our findings show that while CoT prompting helps reduce hallucination frequency, it also tends to obscure critical signals used for detection, impairing the effectiveness of various detection methods. Our study highlights an overlooked trade-off in the use of reasoning. Code is publicly available at: https://github.com/ECNU-Text-Computing/cot-hallu-detect.

1 Introduction

Large language models (LLMs) have demonstrated remarkable performance across a range of natural language processing (NLP) tasks (Qin et al., 2024; Yan et al., 2024; Song et al., 2025), showing a strong ability to follow instructions and answer questions (Zhao et al., 2024; Hadi et al., 2024). However, LLMs may also exhibit *hallucinations*. Specifically, they may produce factually incorrect or semantically irrelevant answers to a given prompt without indicating any lack of confidence (Ji et al., 2023; Yang et al., 2023; Hadi et al., 2024).

To reduce hallucinations in LLMs, extensive research has focused on hallucination detection and mitigation methods (Ji et al., 2023; Huang et al., 2024; Tonmoy et al., 2024). Existing detection methods are built upon various aspects of LLMs, including output consistency (*e.g.*, evaluating the consistency of responses to multiple identical queries) (Wang et al., 2023; Manakul et al., 2023), analysis of internal states (*e.g.*, computing confidence metrics) (Chuang et al., 2024; Chen et al., 2024d; Sriramanan et al., 2024), and self-evaluation of knowledge (*e.g.*, querying its certainty) (Kumar et al., 2024; Diao et al., 2024). Regarding hallucination mitigation, chain-of-thought (CoT) prompting improves response reliability and interpretability by guiding LLMs to generate intermediate reasoning steps before arriving at a final answer (Wei et al., 2022; Kojima et al., 2024). In contrast to more complex solutions such as fine-tuning LLMs or improving decoding strategies, CoT prompting is widely adopted due to its simplicity and effectiveness (Liu et al., 2023; Yang et al., 2024a). The performance improvements shown by GPT-4o (OpenAI, 2024) and DeepSeek-R1 (Guo et al., 2025b) through enhanced reasoning have further fueled growing research interest in reasoning-centric LLMs.

However, as illustrated in Figure 1, the use of CoT prompting may diminish the effectiveness of hallucination detection methods. When an LLM is prompted to perform step-by-step reasoning, it semantically amplifies the LLM's internal confidence in its output. As a result, even when the final answer deviates from the ground truth, the LLM tends to produce incorrect tokens with high confidence (*i.e.*, high probability). Consequently, hallucinations induced by CoT prompting often appear more plausible, making them more difficult for detection methods to identify.

Although several counterexamples indicate that

*Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 1272–1305
November 4-9, 2025

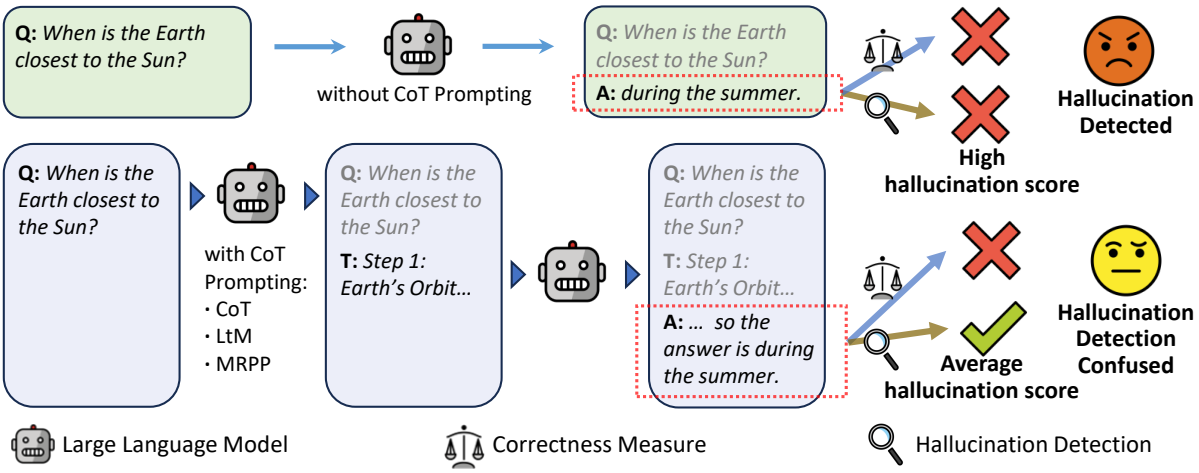

Figure 1: An example of CoT prompting confusing hallucination detection. When the input *When is the Earth closest to the Sun?* is provided to the LLM, it generates a hallucinatory response (*during the summer*) without CoT prompting. In this case, the hallucination detection method assigns a high hallucination score. However, after incorporating CoT prompting, where the LLM introduces reasoning steps (*e.g.*, *Step 1: Earth's Orbit...*), the LLM still produces the same incorrect response, but the detection method assigns only an average hallucination score.

reasoning induced by CoT prompting may negatively impact hallucination detection, the generality of this effect has not been established in prior research. To bridge this gap, we conduct a series of empirical evaluations. We start with pilot experiments on three multiple-choice question answering (MCQA) tasks: CommonSenseQA, ARC-Challenge, and MMLU. In these experiments, we observe that CoT prompting significantly alters the probability distribution of the final answer tokens. Building on these findings, we perform a more extensive set of comparative experiments using two representative types of datasets: (1) complex fact-based question answering datasets, including TriviaQA, PopQA, HaluEval, and TruthfulQA; and (2) a summarization dataset requiring deep contextual understanding, named CNN/Daily Mail. We also incorporate four well-established, recently proposed hallucination detection methods to identify hallucinations. Furthermore, we evaluate three mainstream CoT prompting strategies across both instruction-tuned and reasoning-oriented LLMs, including LLaMA-3.1-8B-Instruct, Mistral-7B-Instruct-v0.3, LLaMA-3.1-70B-Instruct, and DeepSeek-R1-Distill-Llama-8B. To comprehensively assess the impact of CoT prompting on hallucination detection, we examine three key dimensions: (1) shifts in the distribution of hallucination detection output scores; (2) changes in detection accuracy; and (3) variations in the confidence levels of the detection methods.

Our experimental results reveal a dual effect of CoT prompting: while it enhances LLM performance, it simultaneously disrupts hallucination detection mechanisms. This disruption arises because CoT prompting systematically alters the distribution of hallucination scores, leading to a counter-intuitive trade-off. Particularly, improved task performance is accompanied by reduced hallucination detectability. The extent of this effect varies across detection paradigms: consistency-based methods exhibit greater robustness, whereas internal-state-based and self-evaluation-based methods are more susceptible to reasoning. Notably, advanced detection methods lose their typical advantage over perplexity-based baselines under CoT prompting. These findings suggest that reasoning-enhanced inference reshapes the landscape of hallucination detection, making hallucination signals more subtle and harder to capture.

Our contributions are threefold: (1) We conduct a series of experiments, including pilot and comprehensive evaluations, to empirically investigate the impact of CoT-induced reasoning on hallucination detection. (2) We reveal a double-edged effect of CoT prompting: while it enhances LLM performance, it simultaneously obscures common hallucination features. (3) Our results show that internal-state-based and self-evaluation-based methods exhibit reduced robustness under CoT prompting.

2 Preliminary

CoT Prompting Methods. In this study, we focus on three representative methods: Zero-shot

Chain of Thought (**CoT**, Kojima et al. (2024)), Least-to-Most (**LtM**, Zhou et al. (2023)), and Minimum Reasoning Path Prompting (**MRPP**, Chen et al. (2024c)). Further details and formal definitions of these methods are provided in Appendix C.

Unless otherwise specified, all CoT prompting methods are implemented in a *zero-shot* configuration. This ensures that the LLM operates without access to task-specific examples, allowing us to isolate the effects of different reasoning strategies. Following prior work, we design prompt templates tailored to each experimental configuration to align with specific task requirements.

Hallucination Detection Methods. We focus on hallucination detection methods that do not rely on external knowledge bases (*e.g.*, fact-checking (Li et al., 2024)) or require additional fine-tuning and specialized training (*e.g.*, ITI (Li et al., 2023b) and SAPLMA (Azaria and Mitchell, 2023)). This choice is motivated by the limited generalization ability of such methods across tasks and domains. Instead, we examine four main categories of hallucination detection methods: **consistency-based** methods (*e.g.*, SelfCheckGPT-nli (Manakul et al., 2023)); **internal-state-based** methods (*e.g.*, Perplexity, In-Context Sharpness (Chen et al., 2024d), and LLM-Check (Sriramanan et al., 2024)); **self-evaluation-based** methods (*e.g.*, Verbalized Certainty (Kumar et al., 2024)); and **hybrid** methods (*e.g.*, INSIDE (Chen et al., 2024a) and SelfCheckGPT-Prompt (Manakul et al., 2023)), which combine features from multiple categories. For comprehensive descriptions and further details of these methods, see Appendix D.

3 Pilot Experiment and Observation

Recent studies on multiple-choice question answering (MCQA) tasks (Kumar et al., 2024; Chen et al., 2024d; Quevedo et al., 2024) have identified a strong correlation between the probability assigned by LLMs to answer tokens and the correctness of their final predictions. In these datasets, which are primarily composed of factual questions, we adopt the common practice of treating answer incorrectness as a proxy for hallucination labels (Kumar et al., 2024). Motivated by these insights, we perform a pilot experiment to analyze, at the token level, how reasoning with chain-of-thought (CoT) prompting shapes the distribution of probabilities over possible answer tokens. Our analysis pays special attention to instances where CoT prompting fails to eliminate incorrect (*i.e.*, hallucinated) answers.

3.1 Pilot Experimental Setting

Datasets. We use three widely adopted MCQA datasets in our study. **CommonsenseQA** (validation split, 1,221 samples; Talmor et al. (2019)) is a benchmark designed to evaluate commonsense reasoning; we use the validation split since ground-truth answers are unavailable for its test set. **AI2 ARC** (ARC-Challenge, test split, 1,172 samples; Boratko et al. (2018)) is a challenging dataset that focuses on science-related questions requiring advanced reasoning. **MMLU** (test split, 14,042 samples; Hendrycks et al. (2021)) is a large benchmark that spans a broad range of academic and professional domains, providing a comprehensive assessment of general knowledge and reasoning ability.

LLMs. We conduct the pilot experiments using three open-source LLMs, covering both instruction-tuned and reasoning-oriented variants: **Llama-3.1-8B-Instruct** (Dubey et al., 2024), a large-scale model fine-tuned for instruction following; **Mistral-7B-v0.3-Instruct** (Jiang et al., 2023), a compact model trained via instruction tuning; and **DeepSeek-R1-Distill-Llama-8B** (Guo et al., 2025b), a reasoning-oriented model distilled from larger LLMs.

Metrics. We evaluate the LLMs using the following three metrics: **Accuracy**, which measures overall task performance by assessing the LLM's ability to predict the correct option; **Entropy**, which quantifies the uncertainty in the LLM's probability distribution over answer options, reflecting its confidence in decision-making; and **AUROC** (Area Under the Receiver Operating Curve), which treats answer correctness as a proxy hallucination label and evaluates how well the token-level probability distribution distinguishes between hallucinated and non-hallucinated responses.

The implementation details of the pilot experiment are available in Appendix F.

3.2 Observations

The experimental results are shown in Table 1. Overall, LLMs prompted with CoT consistently achieve higher accuracy across all datasets. For example, with Llama-3.1-8B-Instruct, CoT improves accuracy (%) from 90.5, 90.16, and 80.79 to 97.67, 94.96, and 91.57 on the three datasets, respectively. However, the effectiveness of more advanced CoT variants is less consistent. While both LtM and

Table 1: Performance comparison of LLMs under CoT prompting on three token-level MCQA tasks. Experimental results show that CoT prompting improves accuracy and reduces entropy, suggesting increased confidence in the LLM's predictions. However, AUROC scores also decline, indicating that token-level probabilities become less reliable as indicators of hallucinations. Detailed results for Mistral-7B-v0.3-Instruct are provided in the Appendix.

	ARC-Challenge			CommonSenseQA			MMLU		
	Acc	Entropy	AUROC	Acc	Entropy	AUROC	Acc	Entropy	AUROC
DeepSeek-R1-Distill-Llama-8B									
base	67.41	67.70	78.15	59.26	82.56	73.11	54.62	80.74	71.78
LtM	87.36	5.25	69.37	73.52	6.22	69.26	76.92	11.10	68.33
MRPP	88.05	6.93	72.90	72.62	7.13	67.44	76.54	12.08	67.52
CoT	87.94	4.86	70.15	72.38	5.45	68.34	77.36	10.56	64.23
Llama-3.1-8B-Instruct									
base	90.50	23.96	85.98	90.16	26.33	79.75	80.79	46.94	79.97
LtM	93.42	18.14	78.07	89.93	27.82	76.17	85.67	36.31	78.62
MRPP	97.70	6.39	71.81	96.50	9.90	73.15	93.99	15.56	72.54
CoT	97.67	6.63	77.39	94.96	13.50	74.69	91.57	21.39	76.07

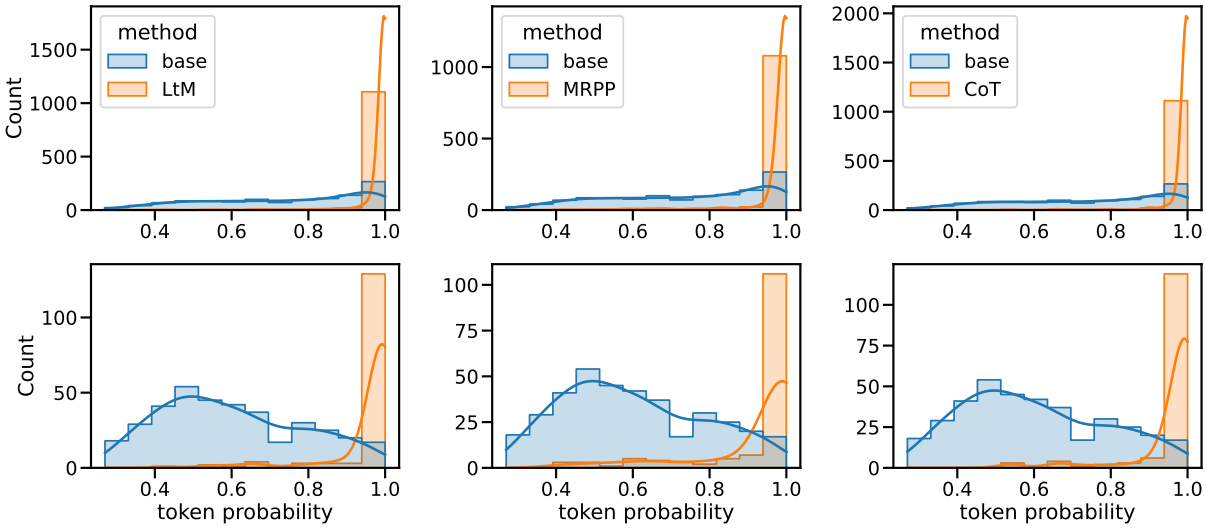

Figure 2: Comparison of token probability distributions for DeepSeek-R1-Distill-Llama-8B on ARC-Challenge. The upper row shows all samples; the lower row shows incorrect predictions after CoT prompting. Kolmogorov-Smirnov test confirms significant differences ($p < 0.01$) between conditions with and without CoT prompting.

MRPP employ multi-path reasoning, MRPP further incorporates confidence-based validation. As a result, LtM may suffer from ambiguity when faced with multiple plausible reasoning paths, achieving better performance than the base Llama but falling short of standard CoT. In contrast, MRPP outperforms even the basic CoT approach.

In addition, after applying CoT prompting, entropy consistently decreases across all LLMs. This suggests that LLMs become more confident in their predictions by assigning higher probabilities to their selected answers. For instance, Llama's entropy drops from 23.96, 26.33, and 46.94 to 6.63, 13.50, and 21.39, respectively, after applying CoT. Similar reductions are observed with other CoT strategies, reflecting a general improvement in model confidence.

The DeepSeek-R1-Distill-Llama-8B model, which is specifically optimized for reasoning abilities, performs substantially below expectations when reasoning is disabled. Its accuracy on the three datasets drops to 67.41, 59.26, and 54.62, respectively. When CoT is enabled, its performance improves significantly, though it still lags behind Llama in terms of accuracy. Interestingly, under CoT prompting, DeepSeek exhibits lower entropy than Llama across all datasets. This suggests that DeepSeek is more confident in its predictions, even though it produces more incorrect answers.

To better illustrate how confident LLMs are when producing hallucinated answers (*i.e.*, incorrect predictions), we compare their token-level probability distributions before and after applying CoT prompting. This comparison is conducted for

both all samples and hallucinated samples. The results are presented in Figure 2. Before CoT prompting, LLMs show relatively uniform probability distributions across samples. For hallucinated samples, the distributions are clearly left-skewed, indicating that incorrect answers tend to receive lower probabilities. In contrast, after CoT prompting, the probability distributions shift such that LLMs assign similarly high probabilities to both correct and incorrect answers. This increased confidence, especially in wrong predictions, underscores the complex influence of CoT prompting on token-level probabilities, ultimately making hallucinations harder to detect.

Previous studies have shown that token probabilities produced by LLMs do not always reliably indicate the presence of hallucinations (Mahaut et al., 2024; Lin et al., 2022a; Chen et al., 2025). Instead, more robust signals can be extracted from the internal states of LLMs to better capture hallucination patterns. To explore this, we conduct a more comprehensive set of experiments using recent hallucination detection models. We investigate whether CoT prompting influences the clarity of hallucination-related signals during the reasoning process.

4 Main Experiments

4.1 Pipeline

This section outlines the workflow used to assess whether CoT prompting interferes with hallucination detection, as illustrated in Figure 3. We begin by determining whether the LLM-generated response constitutes a hallucination based on the ground-truth response. Next, we use a detection model to compute a hallucination score for each output, which is then used to predict whether the output is a hallucination. Finally, we examine the impact of CoT-generated reasoning on hallucination detection from three perspectives. These include four evaluation metrics.

Correctness Measurement. For the responses generated by the LLM, we evaluate their alignment with the ground truth in the dataset based on text matching and semantic consistency. This evaluation determines whether a response constitutes a hallucination.

Hallucination Detection Score. During the response generation process, we apply four mainstream hallucination detection methods to compute the likelihood score of hallucination for each response. Based on these scores, we further predict whether a given response constitutes a hallucination.

Evaluation Dimensions. Based on the hallucination labels and detection scores, we analyze the behavior of hallucination detection methods along three dimensions: score distribution, detection performance, and method confidence. These dimensions range from surface-level statistical patterns to deeper model-level insights. Specifically: (1) changes in the distribution of hallucination scores are evaluated using a distribution consistency test; (2) detection performance is assessed via AUROC and the monotonic relationship; and (3) method confidence is measured using Expected Calibration Error (ECE). Further details are provided in Appendix E.

4.2 Experimental Setup

Datasets. We use four widely adopted QA datasets that require complex reasoning and one summarization dataset. The summarization task demands strong in-context understanding. Specifically, we include TruthfulQA (Lin et al., 2022b), a dataset explicitly designed for hallucination detection. We use its validation subset, which comprises 817 questions. In addition, TriviaQA (Joshi et al., 2017) is a large-scale reading comprehension dataset. We utilize its `rc.wikipedia.nocontext` split and validation subset to support our experiments. PopQA (Mallen et al., 2023) contains questions targeting a long-tail distribution of entities. We use its test subset in a closed-book setting to avoid external knowledge access during evaluation. HaluEval (Li et al., 2023a) evaluates the factual accuracy of LLM-generated content by distinguishing between hallucinatory and accurate outputs. We use its QA subset of non-hallucinatory samples as a benchmark for correctness. For summarization, we employ the CNN/Daily Mail dataset (Hermann et al., 2015), where the LLM is required to generate a summary based on a given news article. This task emphasizes faithful summarization, requiring that generated summaries do not deviate from the source content.

LLMs. Some hallucination detection methods require access to the internal states of LLMs. Thus, we conduct our experiments using open-source LLMs. To ensure diversity across model families, parameter scales, and training objectives, we select four LLMs, including Llama-3.1-8B-Instruct, Mistral-7B-Instruct-v0.3, Llama-3.1-80B-Instruct,

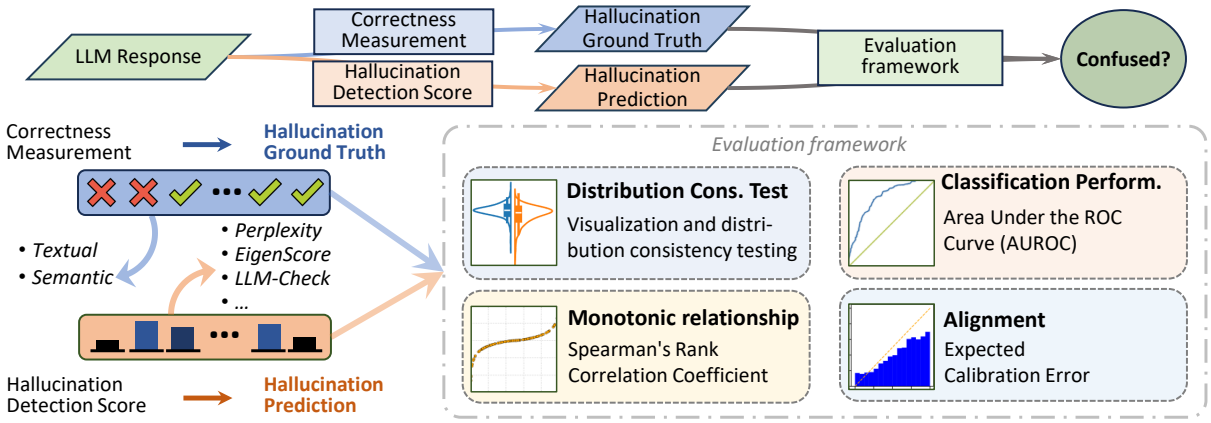

Figure 3: The workflow for evaluating the impact of CoT prompting on hallucination detection. The upper section presents an overview, while the lower section highlights key implementation details. Given the LLM's output, hallucinations are identified by comparing against ground-truth answers. A hallucination detection model then generates a score, which is used to produce predicted labels. Finally, detection performance is evaluated across four dimensions defined in the evaluation framework.

and DeepSeek-R1-Distill-Llama-8B.

Implementation details of the experiments are provided in Appendix F.

5 Results and analysis

This section presents a comprehensive analysis of the experimental results, focusing on three key aspects: (1) the performance enhancement of LLMs through CoT prompting across diverse tasks and datasets; (2) the impact of CoT prompting on the internal mechanisms of LLMs, as revealed by the distribution shifts in hallucination detection scores; and (3) the consequential effects on the performance and reliability of hallucination detection methods. Our findings demonstrate that while CoT prompting significantly improves output quality, it also alters the models' internal representations and confidence calibration. As a result, this ultimately complicates the hallucination detection process.

We validate our automated metrics by having two annotators manually annotate 150 samples, achieving a Cohen's Kappa of 0.863. The Spearman's ρ between our automated scores and human annotations is 0.690 for Exact Match, 0.656 for ROUGE-L F1, and 0.683 for semantic similarity. These results confirm that our automated evaluations reliably reflect human judgment.

5.1 Performance Comparison of CoT Methods Across Different Datasets

As shown in Table 2, the CoT prompting methods significantly enhance the performance of LLMs across various tasks. This is consistent with the results of our pilot experiments. By encouraging LLMs to explicitly generate intermediate reasoning steps, CoT prompting helps activate their latent knowledge and reasoning capabilities, leading to more accurate responses.

Notably, LLM performance rankings vary between the two evaluation approaches, namely textual matching and semantic consistency. This highlights the challenge posed by diverse linguistic expressions in text generation. While evaluation methodology is not the primary focus of this study, we adopt both metrics in subsequent experiments to assess response correctness. This approach helps to mitigate ambiguity, and the resulting correctness serves as the ground truth for hallucination detection.

Furthermore, not all CoT variants consistently improve performance, suggesting that the reasoning process may have nuanced effects on output quality. Interestingly, DeepSeek performs better on summarization tasks, which demand stronger contextual understanding. This observation further supports the notion that incorporating reasoning can enhance LLM performance in such scenarios.

5.2 Comparison of Score Distributions for Hallucination Detection

To examine the impact of CoT prompting on the internal decision-making processes of LLMs, we employ LLM-Check to compare the distributions of Hidden Score and Attention Score with and without CoT prompts. The Hidden Score reflects the LLM's internal confidence in its own output, while

Table 2: Results of CoT prompting influencing the performance of LLMs across various datasets. The table presents results for Llama-3.1-8B-Instruct and DeepSeek-R1-Distill-Llama-8B on HaluEval, PopQA, TriviaQA, and CNN/Daily Mail (C./D.M.). The results indicate that CoT prompting substantially improves LLM performance on these tasks. Please refer to the appendix for additional tasks and LLMs.

	HaluEval	PopQA	TriviaQA	C./D. M.
Llama-3.1-8B-Instruct, Textual				
base	30.26	30.31	79.41	22.13
CoT	36.59	35.05	81.60	23.33
LtM	36.98	33.57	80.02	22.72
MRPP	36.04	33.00	80.46	22.83
Llama-3.1-8B-Instruct, Semantic				
base	39.89	48.58	85.61	91.80
CoT	77.37	65.30	85.17	91.07
LtM	38.86	46.79	85.63	90.79
MRPP	78.10	68.03	85.60	90.82
DeepSeek-R1-Distill-Llama-8B, Textual				
base	19.19	16.09	61.78	20.04
CoT	29.21	22.09	63.16	22.54
LtM	30.15	19.01	64.10	21.39
MRPP	29.21	18.79	64.34	21.68
DeepSeek-R1-Distill-Llama-8B, Semantic				
base	62.30	55.47	72.28	91.26
CoT	78.53	63.15	80.90	92.35
LtM	79.57	64.10	83.34	91.88
MRPP	79.35	64.34	83.02	91.36

the Attention Score captures the degree of focus allocated to different input segments.

We use the Kolmogorov-Smirnov (K-S) test to evaluate whether these score distributions differ significantly between conditions with and without CoT prompting. As shown in Figure 4, both scores exhibit statistically significant shifts ($p < 0.01$), indicating that CoT prompting induces substantial changes in the LLM's internal representations.

For instruction-tuned LLMs such as LLaMA, the scores are relatively dispersed in the absence of CoT prompting, suggesting that hallucination detection methods can effectively differentiate between hallucinated and faithful responses. However, with CoT prompting, the scores become tightly clustered. This implies that the detector may collapse to predicting a single class, thereby losing its discriminative capability.

In contrast, reasoning-oriented LLMs such as DeepSeek exhibit concentrated score distributions even without CoT, suggesting that internal signals are less indicative of hallucination. CoT prompting still significantly alters the distribution, suggesting that explicit reasoning substantially reshapes the LLM's internal state, even when the LLM is already strong in reasoning tasks.

Overall, these findings suggest that CoT prompting not only enhances output quality but also modulates the LLM's internal confidence and attention.

5.3 Performance Comparison of Hallucination Detection Methods

We conduct a comprehensive evaluation of how CoT prompting affects the performance of hallucination detection methods. Our experiments span four benchmark datasets, four LLMs, two hallucination annotation protocols (*i.e.*, Textual or Semantic methods for establishing ground truth), and eight representative detection approaches. For each combination, we evaluate performance under four prompting scenarios: one base scenario (without CoT) and three distinct CoT prompting methods (CoT, LtM, and MRPP). This results in a total of 768 experimental configurations (4 benchmark datasets $\times$ 4 LLMs $\times$ 2 hallucination protocols $\times$ 8 detection methods $\times$ 3 CoT prompting methods). For each configuration, the baseline performance without CoT prompting serves as the point of comparison.

Table 3 summarizes, grouped by LLM and dataset, the number of experimental configurations in which a detection method's AUROC score declined under a CoT prompting scenario compared to the base scenario. Each cell in the table has a maximum value of 12, representing the 4 models evaluated under the 3 CoT methods. Overall, more than half of the configurations (465 out of 768) exhibit performance degradation, suggesting that CoT prompting exerts a broad and substantial negative impact on hallucination detection.

Notably, TruthfulQA is treated separately due to its distinct evaluation protocol. To further assess the generality of our findings, we also evaluate a model-training-based method, MIND (Su et al., 2024). We observe that it likewise suffers degraded hallucination detection performance under CoT prompting (see Appendix G for details).

Our findings reveal that CoT prompting generally impairs the effectiveness of hallucination detection, with the degree of impact varying by method. Self-evaluation-based approaches (*e.g.*, SelfCheckGPT-Prompt and Verbalized Certainty) are most adversely affected. These methods rely heavily on confidence signals present in the LLM's output. However, these signals are easily distorted by CoT-style reasoning. As a result of CoT prompting, LLMs often generate additional justificatory

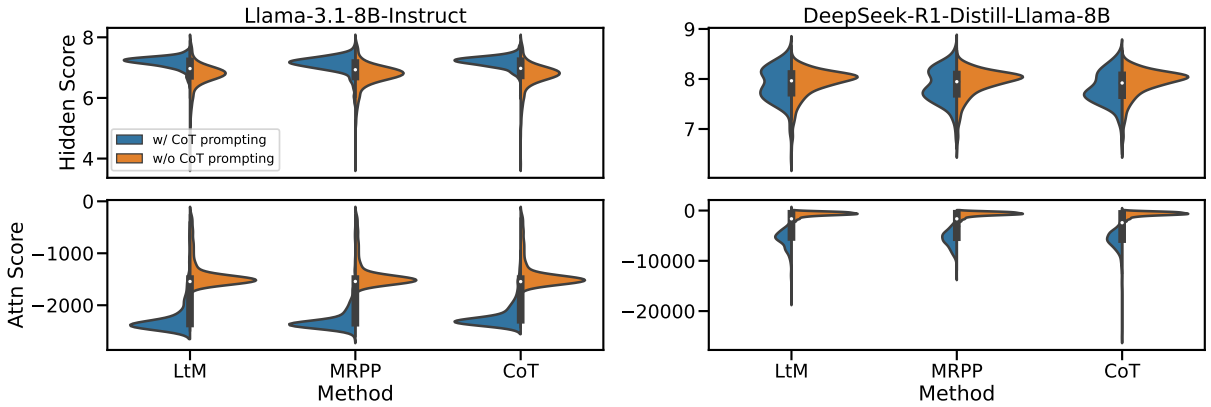

Figure 4: Hidden score and attn score distributions for incorrect predictions. Each plot represents results across different CoT prompting methods (x-axis) and their corresponding scores (y-axis). Significant shifts in score distributions are observed for all hallucination detection methods after incorporating CoT prompting methods, suggesting that CoT prompting methods substantially impact hallucination detection scores: even when hallucinations exist, the scores of hallucination detection methods are biased towards *not hallucination*.

Table 3: Number of experimental configurations in which the AUROC for hallucination detection decreases after applying CoT prompting, relative to the base scenario. Each cell can contain up to 12 configurations, corresponding to 3 CoT variants $\times$ 4 LLMs.

	HaluEval		PopQA		TriviaQA		CNN/Daily Mail	
	Textual	Semantic	Textual	Semantic	Textual	Semantic	Textual	Semantic
PPL	9	6	12	6	10	3	10	8
Sharpness	9	6	12	6	10	4	10	11
Eigen Score	3	8	4	7	9	6	4	7
SelfCk-Prompt	10	7	12	8	9	7	9	6
SelfCk-NLI	9	6	6	2	12	9	8	4
Hidden Score	8	9	4	6	8	5	6	7
Verbalized Certainty	9	7	6	9	11	8	10	7
Attn Score	5	4	0	7	1	5	12	7

content, which can obscure factual inaccuracies. For instance, in response to the TriviaQA question *What was Walter Matthau's first movie?*, an LLM under LtM prompting may include phrases like *Based on reliable sources such as IMDb and Wikipedia....* Although the answer is incorrect, the Verbalized Certainty score may increase from 0.2 to 1.0, illustrating how CoT prompting can mask hallucinations.

By contrast, consistency-based methods (*e.g.*, EigenScore and SelfCheckGPT-NLI) show greater robustness to CoT prompting. These methods leverage multi-sample reasoning to identify inconsistencies, which remain informative even in the presence of CoT prompts. In the above example, consistent sampling occasionally yields the correct answer *The Kentuckian* (EigenScore: 3/15; SelfCheckGPT-NLI: 1/20), albeit infrequently. Although these approaches are often criticized for their high computational overhead, they demonstrate strong resilience and reliability in scenarios that require extensive reasoning.

Another notable observation is that, although advanced hallucination detection methods typically outperform simple perplexity-based baselines under standard conditions, this advantage diminishes when CoT prompting is applied. Without CoT, perplexity-based methods consistently underperform, aligning with prior studies (Chen et al., 2024a; Sriramanan et al., 2024; Chen et al., 2025). However, when CoT prompts are introduced, many detection methods fail to surpass the perplexity baseline, and some perform worse. This finding underscores a critical limitation of current detection approaches: they are not designed to handle the structural and discourse-level shifts induced by explicit reasoning. As a result, their effectiveness at detecting hallucinations is diminished.

TruthfulQA. As shown in Table 4, CoT prompting has inconsistent effects on enhancing both the truthfulness and informativeness of responses in the TruthfulQA task across different scenarios. Among

the 144 experimental settings, 81 exhibit performance degradation (see Appendix G for details). Meanwhile, the negative impact of CoT prompting on hallucination detection methods varies significantly among different LLMs. Specifically, Llama exhibits a more pronounced drop in informativeness, Mistral experiences a greater decline in truthfulness, and DeepSeek shows a decline in both dimensions. This variation likely reflects differences in response generation tendencies among the various LLMs. For example, following CoT prompting, Llama tends to emphasize the reasoning process and frequently includes redundant phrases, such as *Based on the above steps...*, in its answers. Although such content highlights the reasoning chain, it may obscure factual inaccuracies or introduce misleading elements. This can interfere with the effectiveness of hallucination detection methods. To mitigate potential biases introduced by the LLaMA-based judge model, we additionally fine-tuned the evaluation model using Qwen. The evaluation results based on the Qwen judge model (see Appendix G for details) are consistent with our previous findings, confirming the robustness of our conclusions.

Table 4: Number of experiments of TruthfulQA in which the AUROC value decreases. For each experiment, AUROC is calculated by comparing the predicted hallucination detection scores with either informativeness (I.) or truthfulness (T.); each cell shows the number of experiments (three using CoT prompting and one without) where AUROC decreases relative to a baseline. The **first two** columns are results for **Llama**-3.1-8B-Instruct, the **middle two** for **Mistral**-7B-Instruct-v0.3, and the **last two** for **DeepSeek**-R1-Distill-Llama-8B.

	I.	T.	I.	T.	I.	T.
Perplexity	**3**	0	0	**2**	**2**	**3**
Sharpness	**3**	0	0	**2**	**2**	**3**
Eigen Score	**2**	**3**	**1**	**1**	0	**3**
SelfCk-Prompt	**1**	**3**	**1**	0	**2**	**3**
SelfCk-NLI	**1**	**2**	0	**3**	**2**	**3**
Verb. Certainty	**2**	**2**	0	**2**	0	0
Hidden Score	**2**	0	**3**	**3**	**2**	**2**
Attn Score	0	**2**	**3**	**3**	**1**	**3**

Monotonic Relationship. The trends observed in monotonic relationships are consistent with those in classification performance. Comprehensive experimental results are provided in Appendix G.

5.4 Confidence in Hallucination Detection Methods

We further employ Expected Calibration Error (ECE) to evaluate the alignment between the hallucination probabilities predicted by the detection methods and the actual correctness of the hallucination detection. This allows us to assess the calibration of model confidence. The results indicate that CoT prompting leads to an increase in ECE across most experimental settings, suggesting that the confidence of detection methods becomes less reliable. Detailed results are provided in Appendix G.

6 Discussion and Conclusion

In this study, we systematically investigate the impact of CoT prompting on hallucination detection in LLMs. We evaluate several representative detection methods and find that, although CoT prompting improves text-level and semantic accuracy across most datasets, it simultaneously undermines hallucination detection performance. More specifically, CoT prompting results in more concentrated detection score distributions, lower classification accuracy, weaker alignment with ground truth, and poorer confidence calibration. The sensitivity of this impact varies according to the detection method: self-evaluation-based approaches are the most affected, whereas consistency-based methods show greater robustness. We further observe that CoT prompting diminishes and occasionally eliminates the performance advantage of advanced detectors over a simple perplexity baseline.

These findings suggest that CoT prompting can obscure hallucination cues and compromise the effectiveness of existing detection systems, highlighting the need to develop more robust approaches specifically tailored to reasoning-oriented LLMs.

Limitations

Due to computational resource constraints, we are unable to conduct experiments on larger-scale LLMs, which limits the generalizability of our conclusions. Moreover, the limitations of current hallucination detection methods prevented us from evaluating our framework on widely used closed-source LLMs. Future research should incorporate such LLMs to achieve a more comprehensive understanding.

Regarding evaluation metrics, although the use of Exact Match and ROUGE-L to assess response correctness in QA tasks is subject to debate, we

select these metrics because of their widespread adoption, computational efficiency, and effectiveness in measuring surface-level accuracy. Future work may consider more specialized metrics to better capture the nuanced challenges of hallucination detection.

Our findings show that while CoT prompting significantly improves the reasoning abilities of LLMs, it also introduces complexities that undermine existing hallucination detection mechanisms, posing risks in high-stakes applications. This underscores the urgent need to develop more robust hallucination detection methods tailored to reasoning-augmented LLM. Addressing the impact of CoT prompting on hallucination detection is beyond the scope of this study and is left for future work.

Ethics Statement

Our study investigates the impact of CoT reasoning on hallucination detection in LLMs. Importantly, this research does not involve the collection or use of personal data. However, we acknowledge the inherent risks associated with LLMs, including potential biases and misinformation. We are committed to conducting our research responsibly, ensuring transparency in our methodologies, and promoting the ethical use of AI technologies.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (72204087), the Shanghai Planning Office of Philosophy and Social Science Youth Project (2022ETQ001), the Chenguang Program of Shanghai Education Development Foundation and Shanghai Municipal Education Commission (23CGA28), the Shanghai Pujiang Program (23PJC030), Young Elite Scientists Sponsorship Program by CAST (YESS20240562), and the Fundamental Research Funds for the Central Universities, China. We also appreciate the constructive comments from the anonymous reviewers.

References

Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it's lying. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, pages 967 – 976.

Michael Boratko, Harshit Padigela, Divyendra Mikkilineni, Pritish Yuvraj, Rajarshi Das, Andrew McCallum, Maria Chang, Achille Fokoue-Nkoutche, Pavan Kapanipathi, Nicholas Mattei, Ryan Musa, Kartik Talamadupula, and Michael Witbrock. 2018. A systematic classification of knowledge, reasoning, and context within the ARC dataset. In *Proceedings of the Workshop on Machine Reading for Question Answering*, pages 60–70, Melbourne, Australia. Association for Computational Linguistics.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2024a. INSIDE: LLMs' internal states retain the power of hallucination detection. In *The Twelfth International Conference on Learning Representations*, pages 88 – 104.

Kedi Chen, Qin Chen, Jie Zhou, Xinqi Tao, Bowen Ding, Jingwen Xie, Mingchen Xie, Peilong Li, and Zheng Feng. 2025. Enhancing uncertainty modeling with semantic graph for hallucination detection. 39:23586–23594.

Lida Chen, Zujie Liang, Xintao Wang, Jiaqing Liang, Yanghua Xiao, Feng Wei, Jinglei Chen, Zhenghong Hao, Bing Han, and Wei Wang. 2024b. Teaching large language models to express knowledge boundary from their own signals. *Preprint*, arXiv:2406.10881.

Qiguang Chen, Libo Qin, Jiaqi WANG, Jingxuan Zhou, and Wanxiang Che. 2024c. Unlocking the capabilities of thought: A reasoning boundary framework to quantify and optimize chain-of-thought. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Shiqi Chen, Miao Xiong, Junteng Liu, Zhengxuan Wu, Teng Xiao, Siyang Gao, and Junxian He. 2024d. In-context sharpness as alerts: An inner representation perspective for hallucination mitigation. In *Forty-first International Conference on Machine Learning*, pages 7553–7567.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. 2024. Dola: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations*.

Shizhe Diao, Pengcheng Wang, Yong Lin, Rui Pan, Xiang Liu, and Tong Zhang. 2024. Active prompting with chain-of-thought for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1330–1350, Bangkok, Thailand. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Guhao Feng, Bohang Zhang, Yuntian Gu, Haotian Ye, Di He, and Liwei Wang. 2023. Towards revealing the mystery behind chain of thought: A theoretical perspective. In *Thirty-seventh Conference on Neural Information Processing Systems*, pages 70757–70798.

Chunxi Guo, Zhiliang Tian, Jintao Tang, Shasha Li, and Ting Wang. 2025a. Multi-pattern retrieval-augmented framework for text-to-sql with poincaré-skeleton retrieval and meta-instruction reasoning. *Information Processing & Management*, 62(3):103978.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025b. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Muhammad Usman Hadi, Qasem Al Tashi, Abbas Shah, Rizwan Qureshi, Amgad Muneer, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, et al. 2024. Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. *Authorea Preprints*.

Guoxiu He, Xin Song, and Aixin Sun. 2025. Knowledge updating? no more model editing! just selective contextual reasoning. *CoRR*, abs/2503.05212.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.

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 *International Conference on Learning Representations*.

Chen Huang and Guoxiu He. 2025. Text clustering as classification with llms. *Preprint*, arXiv:2410.00927.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2024. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.* Just Accepted.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, Xiaojian Jiang, Jiexin Xu, Li Qiuxia, and Jun Zhao. 2024. Tug-of-war between knowledge: Exploring and resolving knowledge conflicts in retrieval-augmented language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16867–16878, Torino, Italia. ELRA and ICCL.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Geunwoo Kim, Pierre Baldi, and Stephen McAleer. 2023. Language models can solve computer tasks. In *Advances in Neural Information Processing Systems*, volume 36, pages 39648–39677. Curran Associates, Inc.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2024. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, pages 22199–22213, Red Hook, NY, USA. Curran Associates Inc.

Abhishek Kumar, Robert Morabito, Sanzhar Umbet, Jad Kabbara, and Ali Emami. 2024. Confidence under the hood: An investigation into the confidence-probability alignment in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 315–334, Bangkok, Thailand. Association for Computational Linguistics.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023a. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore. Association for Computational Linguistics.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023b. Inference-time intervention: Eliciting truthful answers from a language model. In *Thirty-seventh Conference on Neural Information Processing Systems*, pages 41451–41530.

Miaoran Li, Baolin Peng, Michel Galley, Jianfeng Gao, and Zhu Zhang. 2024. Self-checker: Plug-and-play modules for fact-checking with large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 163–181, Mexico City, Mexico. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022a. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022b. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 158–167, Vancouver, Canada. Association for Computational Linguistics.

Xiangyang Liu, Tianqi Pang, and Chenyou Fan. 2023. Federated prompting and chain-of-thought reasoning for improving llms answering. In *Knowledge Science, Engineering and Management*, pages 3–11, Cham. Springer Nature Switzerland.

Matéo Mahaut, Laura Aina, Paula Czarnowska, Momchil Hardalov, Thomas Müller, and Lluis Marquez. 2024. Factual confidence of LLMs: on reliability and robustness of current estimators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4554–4570, Bangkok, Thailand. Association for Computational Linguistics.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfcheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, page 9004–9017.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Chau Nguyen, Phuong Nguyen, and Le-Minh Nguyen. 2025. Retrieve–revise–refine: A novel framework for retrieval of concise entailing legal article set. *Information Processing & Management*, 62(1):103949.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Augustus Odena. 2022. Show your work: Scratchpads for intermediate computation with language models. In *Deep Learning for Code Workshop*.

OpenAI. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Akshara Prabhakar, Thomas L. Griffiths, and R. Thomas McCoy. 2024. Deciphering the factors influencing the efficacy of chain-of-thought: Probability, memorization, and noisy reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3710–3724, Miami, Florida, USA. Association for Computational Linguistics.

Ben Prystawski, Michael Y. Li, and Noah Goodman. 2023. Why think step by step? reasoning emerges from the locality of experience. In *Thirty-seventh Conference on Neural Information Processing Systems*, pages 70926 – 70947.

Libo Qin, Qiguang Chen, Xiachong Feng, Yang Wu, Yongheng Zhang, Yinghui Li, Min Li, Wanxiang Che, and Philip S. Yu. 2024. Large language models meet nlp: A survey. *Preprint*, arXiv:2405.12819.

Ernesto Quevedo, Jorge Yero, Rachel Koerner, Pablo Rivas, and Tomas Cerny. 2024. Detecting hallucinations in large language model generation: A token probability approach. In *ICAI'24-The 26th Int'l Conf on Artificial Intelligence*.

Jiayu Ren, Chenxi Lin, and Guoxiu He. 2025. *Divide and Conquer: Prompting Large Language Models to Identify Personalities in Long Social Posts via Chunked Voting*. Association for Computing Machinery, New York, NY, USA.

Hithesh Sankararaman, Mohammed Nasheed Yasin, Tanner Sorensen, Alessandro Di Bari, and Andreas Stolcke. 2024. Provenance: A light-weight fact-checker for retrieval augmented LLM generation output. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1305–1313, Miami, Florida, US. Association for Computational Linguistics.

Abulhair Saparov and He He. 2023. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. In *The Eleventh International Conference on Learning Representations*.

Xin Song, Zhikai Xue, Guoxiu He, Jiawei Liu, and Wei Lu. 2025. Interweaving memories of a siamese large language model. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(23):25155–25163.

Zayne Rea Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. 2024. MuSR: Testing the limits of chain-of-thought with multistep soft reasoning. In *The Twelfth International Conference on Learning Representations*.

Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. 2024. LLM-check: Investigating detection of hallucinations in large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, page 915–932.

Kaya Stechly, Karthik Valmeekam, and Subbarao Kambhampati. 2024. Chain of thoughtlessness? an analysis of cot in planning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Weihang Su, Changyue Wang, Qingyao Ai, Yiran Hu, Zhijing Wu, Yujia Zhou, and Yiqun Liu. 2024. Unsupervised real-time hallucination detection based on the internal states of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14379–14391, Bangkok, Thailand. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

S. M Towhidul Islam Tonmoy, S M Mehedi Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *Preprint*, arXiv:2401.01313.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Thirty-seventh Conference on Neural Information Processing Systems*, pages 74952–74965.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. *Preprint*, arXiv:2203.11171.

Yuzhang Wang, Peizhou Yu, and Guoxiu He. 2025. *Silent LLMs: Using LoRA to Enable LLMs to Identify Hate Speech*. Association for Computing Machinery, New York, NY, USA.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, pages 24824–24837.

Wenjia Yan, Bo Hu, Yu li Liu, Changyan Li, and Chuling Song. 2024. Does usage scenario matter? investigating user perceptions, attitude and support for policies towards chatgpt. *Information Processing & Management*, 61(6):103867.

Guang Yang, Yu Zhou, Xiang Chen, Xiangyu Zhang, Terry Yue Zhuo, and Taolue Chen. 2024a. Chain-of-thought in neural code generation: From and for lightweight language models. *IEEE Transactions on Software Engineering*, 50(9):2437–2457.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing Yin, and Xia Hu. 2023. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *Preprint*, arXiv:2304.13712.

Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. 2024b. Alignment for honesty. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Jiacheng Yao, Xin Xu, and Guoxiu He. 2025. Metacognitive symbolic distillation framework for multi-choice machine reading comprehension. *Knowledge-Based Systems*, 312:113130.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik R Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, pages 11809–11822.

Yao Yao, Zuchao Li, and Hai Zhao. 2024. GoT: Effective graph-of-thought reasoning in language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2901–2921, Mexico City, Mexico. Association for Computational Linguistics.

Michihiro Yasunaga, Xinyun Chen, Yujia Li, Panupong Pasupat, Jure Leskovec, Percy Liang, Ed H. Chi, and Denny Zhou. 2024. Large language models as analogical reasoners. In *The Twelfth International Conference on Learning Representations*.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. 2024. How language model hallucinations can snowball. In *Forty-first International Conference on Machine Learning*, pages 59670–59684.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models. *Preprint*, arXiv:2309.01219.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2024. A survey of large language models. *Preprint*, arXiv:2303.18223.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans,

Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*.

Appendix

A Theoretical and Practical Implications

The theoretical significance of this study lies in its exploration of the relationship between CoT prompting and hallucination detection. Our research demonstrates that while CoT prompting methods enhance the interpret-ability and reasoning capabilities of LLMs, they simultaneously confuse the task of detecting hallucinations in these LLMs. This highlights a gap in current research: hallucination detection methods must account for the effects introduced by CoT prompting. From a practical perspective, our work underscores the importance of considering a broader range of factors in hallucination detection. Specifically, we show that detection methods should be capable of handling the subtle differences introduced by CoT prompting. This is crucial for ensuring the safe and reliable deployment of LLMs in real-world applications, especially those requiring high levels of factual accuracy, such as healthcare, legal, and scientific fields.

B Additional Related Work

B.1 Large Language Models and Hallucinations

Large language models demonstrate remarkable generalization capabilities, leading to their widespread adoption across a variety of tasks (Zhao et al., 2024; Qin et al., 2024; Yan et al., 2024; Huang and He, 2025). They exhibit strong performance, especially in question-answering tasks (Zhao et al., 2024; Yang et al., 2023; Hadi et al., 2024). However, despite their effectiveness, LLMs are not infallible, and a significant challenge that persists is the phenomenon of hallucinations (Ji et al., 2023; Hadi et al., 2024; Wang et al., 2025).

Hallucinations in LLMs typically manifest when the model generates fabricated or inaccurate information in an attempt to answer a question (Yang et al., 2024b; Tonmoy et al., 2024). These hallucinations can be classified into three categories: Input-Conflicting, Context-Conflicting, and Fact-Conflicting hallucinations (Zhang et al., 2023). Alternatively, they are often categorized into two broad types: factuality hallucination and faithfulness hallucination (Huang et al., 2024). Factuality hallucinations arise when the LLM's output conflicts with established world knowledge. Specifically, these hallucinations arise when the input

prompt lacks relevant factual information (He et al., 2025). In such cases, the LLM relies on its internal parameters and memory (Mallen et al., 2023), resulting in outputs that contradict factual reality.

In the context of QA tasks, particularly closed-book QA, the issue of Fact-Conflicting or factuality hallucinations is especially pronounced (Jin et al., 2024; Zhang et al., 2023). Several benchmarks are designed to assess these types of hallucinations. Examples include TruthfulQA (Lin et al., 2022b) and HaluEval (Li et al., 2023a), which explicitly focus on hallucinations. Other QA datasets, such as TriviaQA (Joshi et al., 2017) and PopQA (Mallen et al., 2023), were not originally designed for hallucination detection but are still widely used to evaluate factual accuracy (Chen et al., 2024d,a,b).

A variety of methods are proposed to mitigate factuality hallucinations (Tonmoy et al., 2024), including prompt engineering, the integration of external knowledge, changes to decoding strategies, and model training or fine-tuning approaches. However, for users interacting directly with LLMs, prompt-based methods, such as Chain-of-Thought (Wei et al., 2022), provide a more accessible and widely adopted solution. The core idea behind CoT is to guide the LLM in generating reasoning steps before providing a final answer, thereby enhancing response accuracy and reducing hallucinations.

B.2 CoT prompting

Chain-of-Thought (CoT) prompting is a technique designed to enhance the performance of LLMs by guiding them to generate intermediate reasoning steps prior to producing a final answer (Wei et al., 2022). Rather than generating a direct one-step response, CoT prompts the LLM to engage in a structured reasoning process, thereby reducing errors and hallucinations that arise from premature conclusions. By leveraging this step-by-step reasoning approach, CoT demonstrates significant improvements in both accuracy and reliability (Yao et al., 2025).

To reduce the manual effort required to provide examples and to better leverage LLMs' prior knowledge, Self-generated Few-shot CoT is introduced (Yasunaga et al., 2024). This approach allows LLMs to autonomously construct relevant examples and knowledge, enabling adaptive reasoning. Remarkably, studies demonstrate that performance improvements associated with reasoning can occur even without explicitly providing examples (Kojima et al., 2024). By simply appending a phrase such as "let's think step by step" to the prompt, a technique known as Zero-shot CoT, LLMs are prompted to generate reasoning steps. This appended phrase (referred to as a "magic phrase" (Stechly et al., 2024)) is identified as a key trigger for enabling reasoning without explicit demonstrations.

Regardless of whether few-shot examples, self-generated examples, or zero-shot techniques are utilized, the primary objective of these methods is to prompt reasoning prior to delivering a final answer. This foundational idea predates the rise of LLMs (Nye et al., 2022; Ling et al., 2017). Beyond Zero-shot CoT, numerous variants emerge to address specific limitations. For example, Tree-of-Thought (ToT) (Yao et al., 2023) and Graph-of-Thought (GoT) (Yao et al., 2024) propose more sophisticated reasoning structures that facilitate backtracking and synergistic outcomes, thereby addressing the linear constraints of CoT reasoning. Other approaches, such as Least-to-Most Prompting (LtM) (Zhou et al., 2023) and Minimum Reasoning Path Prompting (Chen et al., 2024c), aim to decompose complex problems into simpler sub-problems or ensure that reasoning paths remain as concise and clear as possible.

Collectively, these methods share a common emphasis on generating reasoning steps prior to producing a final answer. For clarity, we refer to them as **CoT prompting**. They demonstrate broad utility, consistently enhancing LLM performance across various tasks and domains (Guo et al., 2025a; Kim et al., 2023; Nguyen et al., 2025). Substantial research examines both the empirical and theoretical effectiveness of CoT prompting (Feng et al., 2023; Chen et al., 2024c; Saparov and He, 2023; Prabhakar et al., 2024; Prystawski et al., 2023), establishing it as a robust strategy for performance improvement.

However, despite their widespread success, CoT prompting faces notable limitations (Sprague et al., 2024). Some studies argue that LLMs do not truly learn the underlying algorithms for solving problems (Stechly et al., 2024) and often fail to faithfully explain the reasoning behind their answers (Turpin et al., 2023). By requiring LLMs to generate additional content, such as reasoning steps, CoT prompting may alter the internal states of the LLM (Feng et al., 2023). This raises an intriguing question: could this state alteration impact tasks that rely on these internal states, such as hallucination detection?

B.3 Hallucination Detection

Hallucinations in LLMs are commonly attributed to a discrepancy between the LLM's internal knowledge and real-world knowledge (Huang et al., 2024; Zhang et al., 2024). Some studies further suggest that hallucinations generated in earlier stages of text generation may propagate, leading to subsequent hallucinations (Zhang et al., 2024; Stechly et al., 2024). This observation aligns with our hypothesis that CoT prompting could influence hallucination detection by altering the LLM's internal states.

Our findings focus on a different aspect: the impact of CoT prompting on hallucination detection. Specifically, even in cases where hallucinations do not result from the propagation of earlier errors, hallucination detection methods may exhibit different performance under CoT prompting compared to standard prompting strategies. This indicates that the additional reasoning steps introduced by CoT prompting may modify the internal states of model in ways that undermine the effectiveness of existing detection methods.

Methods for identifying and mitigating hallucinations can generally be classified based on whether they rely on external data. Among methods that do not rely on external data, three main subcategories are identified: **consistency-based**, **internal-state-based**, and **self-evaluation-based** hallucination detection.

Internal-state-based methods utilize the LLM's internal representations or uncertainty during text generation to distinguish between hallucinated and non-hallucinated content. For instance, LLM-check (Sriramanan et al., 2024) assesses variations in token probabilities during text generation. Other approaches, such as DoLa (Chuang et al., 2024) and In-context Sharpness (Chen et al., 2024d), analyze features extracted from hidden states across different layers or time steps during decoding. These insights are then applied to detect and mitigate hallucinations.

Self-evaluation-based methods are founded on the premise that LLMs are capable of verbally evaluating the content they generate (Yao et al., 2023; Wang et al., 2023; Yao et al., 2024; Kumar et al., 2024; Diao et al., 2024). These methods employ prompts to extract the LLM's self-evaluated confidence in its outputs. By treating the LLM's verbalized certainty as an indicator of correctness, these methods identify hallucinations using confidence scores.

Consistency-based methods, such as SelfCheckGPT, operate under the assumption that if an LLM has accurate knowledge and confidence in its answers, it will generate semantically consistent outputs across multiple sampling attempts (Wang et al., 2023; Manakul et al., 2023; Ren et al., 2025). This approach is particularly relevant in factual QA tasks, where semantic consistency among sampled answers serves as an indicator of factuality. Inconsistent outputs, on the other hand, suggest the presence of hallucinations.

These three approaches are not mutually exclusive and can be combined. For example, INSIDE (Chen et al., 2024a) integrates internal-state-based and consistency-based methods, while SelfCheckGPT (Manakul et al., 2023) incorporates both self-evaluation-based and consistency-based techniques.

In contrast, hallucination detection methods that depend on external data are further categorized into retrieval-based and model-training-based approaches. Retrieval-based methods utilize external knowledge bases to validate generated outputs (Min et al., 2023; Sankararaman et al., 2024; Li et al., 2024). However, these methods often face limitations, including reliance on the quality and scope of external data, as well as the complexity of integrating retrieval and comparison systems. Model-training-based methods, such as SAPLMA (Azaria and Mitchell, 2023) and Inference-Time Intervention (ITI) (Li et al., 2023b), necessitate additional labeled data to train auxiliary models for hallucination detection. These approaches, while effective, are resource-intensive and often lack generalization ability. For instance, ITI not only identifies hallucinations but also employs trained probes to address them. Nevertheless, the high resource demands and restricted plug-and-play functionality render these methods less practical for direct application to existing LLMs.

In light of these challenges, our study concentrates on hallucination detection methods that do not rely on external data, with a particular emphasis on consistency-based, internal-state-based, and self-evaluation-based approaches. These methods provide enhanced flexibility and can be seamlessly integrated into existing LLM architectures without incurring additional resource overhead.

C Descriptions of CoT prompting

CoT encourages LLMs to think through problems step by step, fostering logical consistency and improving performance in complex tasks. In our experiments, we use the following adapted template: `Think about it step by step.`

LtM guides the LLM to decompose complex questions into simpler subproblems, solve them individually, and recombine the answers to address the original question. The template used in our study is: `What subproblems must be solved before answering the inquiry?`

MRPP aims to minimize cognitive load by ensuring that each reasoning step is as simple and error-free as possible. The adapted template we employ is: `You need to perform multi-step reasoning, with each step carrying out as many basic operations as possible.`

Given these challenges, our study focuses on hallucination detection methods that do not require external data, particularly consistency-based, internal-state-based, and self-evaluation-based approaches. These methods offer greater flexibility and can be readily applied to existing LLM architectures without additional resource overhead.

D Descriptions of Hallucination Detection Methods

Perplexity serves as a metric for assessing the predictive accuracy of an LLM on a given sequence. It is defined as:

$$PPL(x) = 2^{-\frac{1}{n}\sum_{i=1}^{n}\log P(x_i|x_{<i})} \tag{1}$$

where $x = \{x_1, x_2, \ldots, x_n\}$ denotes the token sequence, and $P(x_i|x_{<i})$ indicates the conditional probability of token x_i given the preceding tokens.

In a well-trained language model, sequences that conform to its learned language patterns are assigned higher probabilities, thereby yielding lower perplexity. By contrast, hallucinated content frequently diverges from the statistical patterns present in the training data, resulting in lower token probabilities and higher perplexity. This renders perplexity a straightforward yet effective indicator for hallucination detection.

Sharpness (Chen et al., 2024d) represents an internal-state-based method designed to assess the concentration of a LLM's hidden state during next-token prediction. The method introduces a metric termed *contextual entropy*, which measures the dispersion of the LLM's internal activations across input context tokens. For a predicted token v_p, the contextual entropy is computed as:

$$E(v_p, v_{1:t}) = -\sum_{i=1}^{t} \tilde{P}(activation = i \mid v_p, v_{1:t}) \times \log \tilde{P}(activation = i \mid v_p, v_{1:t}). \tag{2}$$

Here, $\tilde{P}(\text{activation} = i \mid v_p, v_{1:t})$ denotes the normalized activation probability of token v_p relative to each context token v_i. Lower entropy values signify more concentrated activations, implying sharper internal representations and greater confidence in token predictions.

In its original formulation, contextual entropy serves to adjust the original token probability distribution via a weighting coefficient. This adjustment enables the method to regulate the influence of entropy on token probabilities, thereby improving its capacity to detect factually incorrect predictions. By incorporating this entropy-based adjustment, the method successfully identifies discrepancies between high-confidence predictions and factually grounded content, rendering it well-suited for hallucination detection.

Verbalized Certainty (Kumar et al., 2024) is a self-evaluation-based method that enables the LLM to explicitly express its confidence level in the accuracy of its outputs. This is achieved by generating responses to a new query. The method employs prompts to elicit the LLM's self-evaluation of response confidence, utilizing techniques such as *Confidence Querying Prompts*, *Simulation of a Third-Person Perspective*, and *Likert Scale Utilization*.

INSIDE (Chen et al., 2024a) represents a hybrid method that integrates internal-state-based and consistency-based approaches. It assesses self-consistency across multiple responses by calculating the **EigenScore**, which measures semantic diversity in the embedding space. For K generated sequences, the covariance matrix of their sentence embeddings is computed as:

$$\mathbf{\Sigma} = \mathbf{Z}^\top \cdot \mathbf{J}_d \cdot \mathbf{Z} \tag{3}$$

where $\mathbf{\Sigma} \in \mathbb{R}^{K\times K}$ is the covariance matrix, $\mathbf{Z} = [\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_K] \in \mathbb{R}^{d\times K}$ represents the embedding matrix, and $\mathbf{J}_d = \boldsymbol{I}_d - \frac{1}{d}\mathbf{1}_d\mathbf{1}_d^\top$ is the centering matrix. The EigenScore is defined as the log-determinant of the regularized covariance ma-

trix:

$$E(\mathcal{Y}|\boldsymbol{x},\boldsymbol{\theta}) = \frac{1}{K}\log\det(\boldsymbol{\Sigma} + \alpha \cdot \mathbf{I}_K) \quad (4)$$

where $\alpha \cdot \mathbf{I}_K$ ensures numerical stability, and the eigenvalues λ_i are utilized to compute:

$$E(\mathcal{Y}|\boldsymbol{x},\boldsymbol{\theta}) = \frac{1}{K}\sum_{i=1}^{K}\log(\lambda_i). \quad (5)$$

This metric effectively captures the self-consistency of generated responses and proves particularly adept at identifying hallucinations by detecting inconsistencies across multiple samples.

SelfCheckGPT (Manakul et al., 2023) utilizes consistency-based methods, including Natural Language Inference (NLI) and self-evaluation-based techniques, to evaluate hallucinations. In its NLI-based form, SelfCheckGPT compares LLM-generated responses with stochastically generated alternatives. It then categorizes the relationships between the hypothesis (response) and premise (context) into three categories: *entailment*, *neutral*, or *contradiction*. The contradiction score serves as the SelfCheckGPT score. Alternatively, in its self-evaluation-based form, SelfCheckGPT directly queries the LLM to verify whether specific sentences are supported by the generated context through carefully designed prompts.

LLM-Check (Sriramanan et al., 2024) represents an internal-state-based method that assesses hallucinations by analyzing the covariance structure of token embeddings within a single response. Similar to INSIDE's EigenScore, LLM-Check computes the covariance matrix of hidden representations:

$$\boldsymbol{\Sigma}_2 = \mathbf{H}^\top\mathbf{H} \quad (6)$$

and defines the hallucination score as the mean log-determinant of $\boldsymbol{\Sigma}_2$:

$$\text{Score} = \frac{2}{m}\sum_{i=1}^{m}\log\sigma_i \quad (7)$$

where σ_i are the singular values of $\mathbf{H}$.

E Details on the Evaluation Framework

The evaluation framework aims to systematically examine the relationship between hallucination scores and correctness under both CoT prompting and non-CoT prompting conditions. The framework includes the following components:

Distribution Consistency Test. We use statistical methods, such as the Kolmogorov-Smirnov test, to compare the distribution of hallucination scores with and without CoT prompting. This analysis determines whether the introduction of CoT prompting significantly alters the distribution of hallucination scores. It also provides an intuitive measure of how CoT prompting influences hallucination detection.

Classification Performance. We treat correctness as the ground truth (binary: correct or incorrect) and hallucination scores as predictions. We then use AUROC to quantify the performance of hallucination detection methods before and after applying CoT prompting. This metric quantifies the degree to which CoT prompting influences the effectiveness of hallucination detection methods.

Monotonic Relationship. Using Spearman's rank correlation coefficient, we evaluate the monotonic relationship between hallucination scores and correctness values. Here, correctness is treated as a continuous variable rather than a binary label. This metric shows whether hallucination detection scores are consistent with correctness under CoT prompting.

Alignment. We use Expected Calibration Error (ECE) to assess how well hallucination scores align with correctness. By treating correctness as the ground truth and hallucination scores as predictions, this metric evaluates whether hallucination detection methods effectively reflect the degree of correctness. This approach goes beyond simply distinguishing between correct and incorrect responses. This is particularly important as some answers may not be strictly correct or incorrect, and well-calibrated hallucination scores should reflect the nuances of correctness.

This evaluation framework collectively enables a comprehensive analysis of whether CoT prompting confuses hallucination detection methods. By observing performance degradation, changes in distribution, and alignment issues, our methodology provides a systematic approach to understanding and quantifying the impact of reasoning steps on hallucination detection.

F Experiments Implementation Details

F.1 Pilot Experiment

The normalization process for token probabilities in candidate options can be mathematically described

as follows:

$$p_{\mathrm{o}} = \frac{p_{\mathrm{t}}[t_{\mathrm{o}}]}{\sum_{i=1}^{N} p_{\mathrm{t}}[t_i]}, \hat{y} = \arg\max_i p_{\mathrm{o}}[i], \quad (8)$$

where t_{o} and p_{o} denote the tokens and normalized probabilities of each candidate option, respectively. N is the total number of candidate options, and $\hat{y}$ is the LLM's predicted option, determined by selecting the one with the highest probability. The predicted answer $\hat{y}$ is then compared to the ground-truth answer to determine whether the LLM's response is correct. A correct prediction implies no hallucination, while an incorrect one corresponds to a potential hallucination.

For experiments involving CoT prompting, we first prompt the LLM to generate reasoning steps and then append a final query about the candidate options to determine the final answer. For the baseline (direct answer) setup, we directly prompt the LLM for the final answer without generating reasoning steps. We generate reasoning steps using the `vLLM`[1] with `temperature` = 0.5 and `max_tokens` = 512.

Once the complete prompt is constructed, we feed it into the LLM to compute token-level probabilities for the candidate options. These probabilities are recorded for subsequent calculations of accuracy, entropy, and AUROC.

By systematically comparing LLMs' performance with and without CoT prompting, we aim to evaluate their ability to improve confidence and accuracy, as well as their impact on hallucination detection.

F.2 Main Experiment

We conduct all experiments using four open-source instruction-tuned LLMs: Llama-3.1-8B-Instruct[2]: A lightweight yet capable language model fine-tuned for instruction-following tasks, balancing efficiency and performance in resource-constrained environments (Dubey et al., 2024). Mistral-7B-v0.3-Instruct[3]: A compact model optimized for competitive reasoning and low-latency inference, trained via advanced instruction tuning (Jiang et al., 2023). Llama-3.1-70B-Instruct[4]: A 70-billion-parameter Transformer-based model designed for high-precision instruction alignment, excelling in complex reasoning and human-preference tasks across diverse domains (Dubey et al., 2024). DeepSeek-R1-Distill-Llama-8B[5]: An 8-billion-parameter distilled variant leveraging knowledge transfer from the 671B DeepSeek-R1 model, specialized in step-by-step mathematical reasoning and efficient deployment on consumer-grade hardware (Guo et al., 2025b).

Unless otherwise specified, all generative processes are conducted using the `vLLM` inference framework with a temperature setting of $T = 0.5$. The system prompt used for response generation was: `You are a helpful assistant.` For the reasoning-oriented model, we first force the output of `<think>\n\n</think>` and then let the model continue generating, thereby implementing the "without CoT" setting.

When generating reasoning content, we apply the corresponding prompts discussed in Section 2. For consistency-based hallucination detection methods such as INSIDE and SelfCheckGPT, which involve stochastic sampling, we adopt the generation parameters reported in their respective papers. For INSIDE, we set `n=15`, `temperature=0.5`, `top_p=0.99`, and `top_k=5`. For SelfCheckGPT, the parameters are `n=20` and `temperature=1`.

For internal-state-based hallucination detection methods, including INSIDE, Sharpness, and LLM-Check, we determine layer-specific settings based on the best-performing configurations reported in the literature. Specifically, for INSIDE, we use layer 17; for Sharpness, layer 26; and for LLM-Check, we use layer 15 for the Hidden Score and layer 23 for the Attention Score. For Sharpness, which involves adjusting the entropy's impact on the original token distribution using a parameter α, we set $\alpha = 1.0$ according to the settings in its original paper.

TruthfulQA offers a training set with 6.9k examples, which we use to fine-tune the `GPT-3-6.7B` model for evaluation purposes. Since OpenAI's API no longer provides access to this model, we instead fine-tune the `Llama-3.1-8B`[6] model using LoRA (Hu et al., 2022) to compute the Truthfulness and Informativeness scores. We conduct fine-

[1] https://docs.vllm.ai/en/v0.6.1.post1/

[2] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[3] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

[4] https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct

[5] https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Llama-8B

[6] https://huggingface.co/meta-llama/Llama-3.1-8B

tuning on the `LLaMA-Factory`[7] platform with the following hyperparameters: `learning rate = 5.0e-05`, `batch size = 32`, `epochs = 5`, `lora_rank = 8`, and `lora_alpha = 16`.

The evaluation metrics are defined in Section 4.1. For Expected Calibration Error (ECE), we apply a normalization process to ensure predictions fall within the $[0, 1]$ range. Specifically, we scale the data within the $3\% - 97\%$ range to $[0, 1]$, while outliers beyond this range are clipped to 0 or 1. This normalization avoids distortions caused by extreme values. Verbalized Certainty is excluded from this analysis because its hallucination detection scores are confined to five distinct levels.

G Results

G.1 MIND Evaluation Details

To explore whether our observations extend to model-training-based detection paradigms, we conducted additional experiments using the MIND framework (Su et al., 2024).

MIND is an unsupervised, internal-state-based hallucination detection framework. Its key advantages include:

- Leveraging pseudo-labeled data automatically constructed from Wikipedia for training, thus avoiding manual annotation.
- Training lightweight classifiers on the LLM's hidden states for real-time detection.

However, it also shares limitations with other model-training-based approaches:

- It requires training an auxiliary detector tailored to each specific target LLM.
- Its performance is dependent on the design and quality of the pseudo-labeled data.

This need for model-specific training is why we primarily focused on non-training-based detection methods in our main experiments.

Experimental Setup. Due to the tight coupling between MIND's released code and its original LLM architecture, our evaluation was limited to the **LLaMA-2-7b-hf** model. We evaluated its performance on the **HaluEval** dataset under different prompting scenarios: standard prompting (*base*), and the three CoT prompting methods studied in this work.

[7] https://github.com/hiyouga/LLaMA-Factory

Results. The results, presented in Table 5, show that MIND's detection performance also degrades when the LLM responds after CoT prompting. This degradation is consistent across both *Textual* and *Semantic* hallucination annotations and across both AUROC and Correlation metrics, reinforcing the broad and challenging nature of the phenomenon we identify.

Table 5: Performance of the MIND detector on HaluEval under different prompting scenarios.

	Textual		Semantic	
	AUROC	Corr	AUROC	Corr
base	0.6765	0.4733	0.5904	0.2941
CoT	0.6521	0.4261	0.5572	0.2307
MRPP	0.6513	0.4304	0.5608	0.2381
LtM	0.6379	0.3983	0.5573	0.2246

G.2 Detail of Results

Table 6: Results of different LLMs and CoT prompting methods on the TruthfulQA dataset (values ×100). All metrics are positively oriented. Improvements after CoT prompting are bolded. Notably, the Reasoning model shows decreased Truthfulness but increased Informativeness after reasoning.

	Truthfulness	Informativeness
Llama-3.1-8B-Instruct		
base	65.73	94.37
CoT	**65.85**	**97.18**
LtM	**74.17**	**97.55**
MRPP	52.63	**95.23**
Mistral-7B-Instruct-v0.3		
base	85.68	93.64
CoT	**87.88**	**97.06**
LtM	80.29	**95.72**
MRPP	83.23	**96.21**
DeepSeek-R1-Distill-Llama-8B		
base	56.18	94.86
CoT	21.18	**97.92**
LtM	25.83	**96.45**
MRPP	22.77	**96.82**

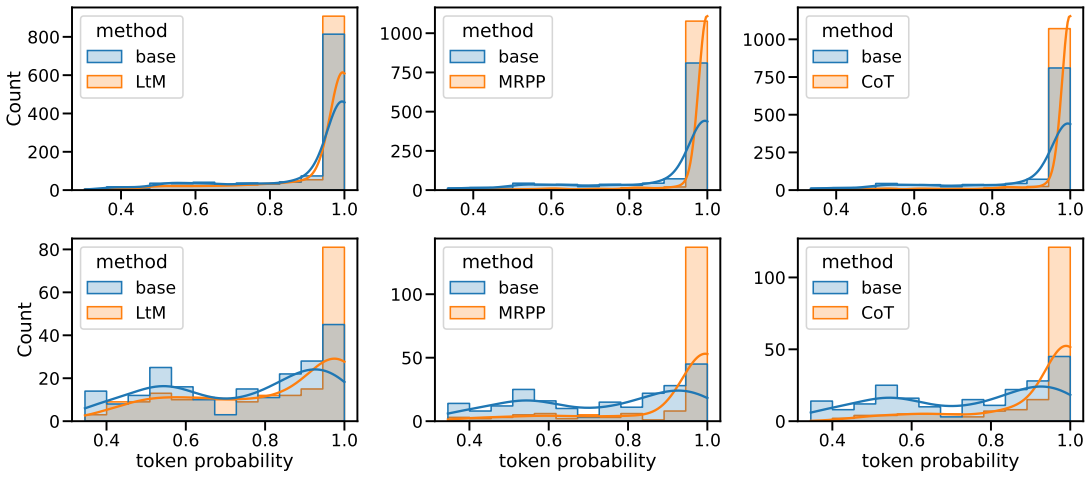

Figure 5: Comparison of token probability distributions for `Llama-3.1-8B-Instruct` on ARC-Challenge. The upper row shows all samples; the lower row shows incorrect predictions after CoT prompting. Kolmogorov-Smirnov test confirms significant differences ($p < 0.01$) between conditions with and without CoT prompting.

Table 7: AUROC results of `Llama-3.1-8B-Instruct` across different datasets. Bolded values indicate cases with reduced performance.

		PPL	Sharpness	Eigen Score	SelfCk -Prompt	SelfCk -NLI	Verbalized Certainty	Hidden Score	Attn Score
Textual HaluEval	base	65.77	65.67	60.75	69.58	62.34	75.37	53.07	58.65
	CoT	67.89	67.42	64.33	**61.93**	69.14	**62.47**	54.51	**53.79**
	LtM	67.77	67.59	63.04	**62.75**	67.79	**62.91**	**51.62**	**56.10**
	MRPP	67.48	67.22	66.73	**62.82**	67.75	**61.54**	57.34	**56.46**
Semantic HaluEval	base	59.76	59.62	62.44	58.34	55.68	65.08	62.86	57.01
	CoT	64.12	64.95	**61.33**	**53.78**	**50.81**	69.59	**61.54**	**54.14**
	LtM	61.14	61.26	**61.92**	**54.29**	58.92	**59.71**	**56.84**	57.97
	MRPP	70.33	69.63	64.20	**51.10**	62.77	77.51	**55.96**	61.79
Textual PopQA	base	75.35	75.38	70.76	65.25	62.90	76.79	66.46	57.58
	CoT	**72.85**	**72.55**	73.02	**54.36**	69.40	**65.08**	**63.43**	60.62
	LtM	**71.96**	**71.79**	70.98	**58.65**	69.57	**66.14**	**61.58**	65.13
	MRPP	**68.73**	**68.53**	**69.91**	**60.17**	68.56	**62.29**	**63.52**	61.29
Semantic PopQA	base	71.90	72.13	72.16	53.05	51.68	80.03	72.07	66.24
	CoT	**66.06**	**66.11**	**66.49**	54.77	56.23	**64.90**	**61.96**	**57.28**
	LtM	73.60	73.62	72.58	56.70	71.80	**67.08**	**64.15**	67.50
	MRPP	**70.99**	**71.10**	**70.32**	**50.64**	64.30	**62.69**	**65.43**	**61.50**
Textual TriviaQA	base	75.56	75.30	70.39	77.20	82.21	64.39	59.43	55.13
	CoT	**69.86**	**69.53**	**67.62**	**69.32**	**77.87**	**59.39**	**57.24**	56.86
	LtM	**70.46**	**70.34**	**68.07**	**66.64**	**71.61**	**58.60**	**58.77**	61.49
	MRPP	**65.67**	**65.52**	**67.55**	**66.40**	**68.85**	**58.38**	**59.19**	61.19
Semantic TriviaQA	base	79.25	79.22	75.27	59.38	65.61	58.84	68.50	67.35
	CoT	**75.03**	**75.16**	**67.52**	**52.97**	**58.69**	**54.75**	**65.92**	**60.65**
	LtM	**75.12**	**75.00**	**69.38**	**55.76**	**62.33**	**55.11**	**61.43**	**61.19**
	MRPP	79.46	**79.18**	**74.27**	63.69	**63.86**	**56.95**	**67.12**	**60.78**

Table 8: AUROC results of `Mistral-7B-Instruct-v0.3` across different datasets. The bolded values indicate cases with reduced performance.

		PPL	Sharpness	Eigen Score	SelfCk -Prompt	SelfCk -NLI	Verbalized Certainty	Hidden Score	Attn Score
Textual HaluEval	base	65.90	65.89	63.45	76.71	67.24	57.13	56.49	58.96
	CoT	**64.96**	**64.96**	64.24	**70.47**	**63.12**	59.08	**55.90**	60.55
	LtM	**63.94**	**63.94**	64.69	**68.64**	**63.24**	61.30	**56.27**	59.73
	MRPP	**64.18**	**64.18**	64.96	**69.11**	**59.26**	62.46	59.47	61.39
Semantic HaluEval	base	63.41	63.41	63.49	59.39	58.04	63.39	57.88	60.93
	CoT	75.51	75.56	**63.05**	**52.94**	**51.29**	75.94	**50.65**	65.49
	LtM	72.70	72.60	79.83	**58.57**	**56.60**	80.34	**52.94**	63.20
	MRPP	76.65	76.74	73.10	**51.89**	63.33	76.63	**51.81**	69.32
Textual PopQA	base	68.01	68.01	62.75	77.25	59.58	64.79	54.91	57.64
	CoT	**62.71**	**62.71**	66.30	**72.28**	**58.96**	**62.31**	61.48	63.67
	LtM	**66.19**	**66.20**	67.54	**71.82**	63.76	**61.90**	63.56	65.50
	MRPP	**60.99**	**61.00**	68.64	**72.11**	**55.24**	67.01	65.51	67.00
Semantic PopQA	base	59.55	59.56	56.97	65.23	60.73	59.62	58.84	59.87
	CoT	**58.47**	**58.47**	**54.32**	**54.29**	**60.26**	**58.30**	**53.42**	**51.63**
	LtM	63.73	63.73	63.43	**60.68**	64.21	**55.92**	61.86	63.93
	MRPP	61.44	61.44	**52.08**	**57.96**	62.96	**53.06**	**55.68**	**55.28**
Textual TriviaQA	base	64.17	64.20	60.37	84.50	73.47	59.00	55.29	57.75
	CoT	66.57	66.81	67.77	**76.36**	**68.47**	**57.95**	60.53	62.07
	LtM	**62.56**	**62.65**	67.28	**74.20**	**68.51**	**58.40**	62.90	63.45
	MRPP	66.38	66.24	69.14	**76.09**	**64.60**	62.13	68.02	67.78
Semantic TriviaQA	base	63.57	63.80	59.24	58.61	61.16	56.49	55.69	57.84
	CoT	**63.00**	**62.94**	64.45	**57.38**	**61.00**	**55.23**	62.85	**55.48**
	LtM	72.94	72.89	77.06	63.66	70.63	57.86	71.43	71.00
	MRPP	71.62	71.61	71.45	65.30	67.96	61.09	65.55	61.25

Table 9: AUROC results of `DeepSeek-R1-Distill-Llama-8B` across different datasets. The bolded values indicate cases with reduced performance.

		PPL	Sharpness	Eigen Score	SelfCk -Prompt	SelfCk -NLI	Hidden Score	Verbalized Certainty	Attn Score
Textual HaluEval	base	0.5287	0.5281	0.6097	0.6464	0.6543	0.6094	0.6094	0.6346
	CoT	**0.5142**	**0.5161**	**0.5192**	**0.6450**	**0.5932**	**0.5492**	**0.5492**	**0.6156**
	LtM	**0.5063**	**0.5089**	**0.5202**	**0.6276**	**0.5895**	**0.5117**	**0.5117**	0.6502
	MRPP	**0.5135**	**0.5129**	**0.5274**	**0.6313**	**0.5773**	**0.5338**	**0.5338**	**0.6258**
Semantic HaluEval	base	0.6648	0.6643	0.6543	0.5033	0.5207	0.5382	0.5382	0.7279
	CoT	**0.6546**	**0.6541**	**0.5513**	**0.5003**	0.5493	**0.5061**	**0.5061**	**0.5058**
	LtM	**0.6571**	**0.6563**	**0.5830**	0.5491	0.5434	**0.5019**	**0.5019**	**0.5136**
	MRPP	**0.6295**	**0.6323**	**0.5610**	0.5389	0.5472	**0.5073**	**0.5073**	**0.5079**
Textual PopQA	base	0.5297	0.5288	0.5387	0.7073	0.6331	0.5134	0.5134	0.5195
	CoT	**0.5096**	**0.5066**	**0.5372**	**0.6752**	**0.5874**	0.5386	0.5386	0.6479
	LtM	**0.5045**	**0.5051**	0.5504	**0.5869**	**0.5302**	0.5647	0.5647	0.5462
	MRPP	**0.5015**	**0.5055**	**0.5290**	**0.5659**	**0.5245**	0.5507	0.5507	0.5403
Semantic PopQA	base	0.6584	0.6587	0.6052	0.5721	0.5124	0.5107	0.5107	0.5606
	CoT	0.6410	0.6380	0.6039	**0.5019**	0.5988	0.6202	0.6202	**0.5179**
	LtM	**0.6010**	**0.5928**	**0.5547**	**0.5189**	0.5768	0.5733	0.5733	**0.5319**
	MRPP	**0.5770**	**0.5744**	**0.5457**	**0.5079**	0.5514	0.5598	0.5598	**0.5300**
Textual TriviaQA	base	0.5417	0.5412	0.5716	0.7345	0.7132	0.5357	0.5357	0.5315
	CoT	**0.5142**	**0.5125**	**0.5257**	**0.6255**	**0.5414**	0.5362	0.5362	0.6504
	LtM	**0.5179**	**0.5083**	**0.5178**	**0.6134**	**0.5565**	**0.5152**	**0.5152**	0.6684
	MRPP	**0.5244**	**0.5218**	**0.5150**	**0.6263**	**0.5431**	**0.5167**	**0.5167**	0.6592
Semantic TriviaQA	base	0.5159	0.5155	0.5261	0.5602	0.6022	0.5857	0.5857	0.5160
	CoT	0.6588	0.6530	0.6027	**0.5163**	**0.5337**	**0.5707**	**0.5707**	0.5352
	LtM	0.6946	0.6915	0.6295	**0.5175**	**0.5822**	0.5937	0.5937	0.5891
	MRPP	0.7004	0.7002	0.6277	**0.5286**	**0.5832**	0.6171	0.6171	0.6046
Textual CNN/Daily Mail	base	0.6709	0.6697	0.5414	0.5077	0.5085	0.6315	0.6315	0.6648
	CoT	**0.6133**	**0.6082**	0.5757	0.5123	**0.5035**	**0.5274**	**0.5274**	**0.6043**
	LtM	**0.6151**	**0.6079**	0.5758	**0.5002**	0.5274	**0.5576**	**0.5576**	**0.6380**
	MRPP	**0.6145**	**0.6082**	0.5555	**0.5068**	**0.5063**	**0.5428**	**0.5428**	**0.6256**
Semantic CNN/Daily Mail	base	0.6769	0.6751	0.5557	0.5258	0.5210	0.5002	0.5002	0.6716
	CoT	0.5823	**0.5734**	0.5587	**0.5075**	0.5298	0.5143	0.5143	0.6942
	LtM	0.5919	**0.5795**	0.5610	**0.5027**	0.5457	0.5231	0.5231	0.6877
	MRPP	0.6004	**0.5901**	0.5798	**0.5001**	0.5242	0.5253	0.5253	0.6998

Table 10: AUROC results of `Llama-3.1-70B-Instruct` across different datasets. Bolded values indicate cases with reduced performance.

		PPL	Sharpness	Eigen Score	SelfCk -Prompt	SelfCk -NLI	Verbalized Certainty	Hidden Score	Attn Score
Textual HaluEval	base	0.7561	0.7561	0.7311	0.6877	0.6923	0.6182	0.7303	0.5721
	CoT	**0.6817**	**0.6802**	0.7694	0.7014	**0.6018**	**0.5503**	**0.6390**	0.7005
	LtM	**0.7223**	**0.7226**	0.7423	**0.6503**	**0.6232**	**0.5542**	**0.6725**	0.6847
	MRPP	**0.7491**	**0.7493**	0.7762	0.6877	**0.6250**	**0.5527**	0.7451	0.6724
Semantic HaluEval	base	0.7241	0.7234	0.6998	0.5782	0.5997	0.5938	0.6019	0.5165
	CoT	**0.6870**	**0.6870**	**0.6527**	0.6226	**0.5514**	**0.5275**	0.6782	0.5889
	LtM	**0.7130**	**0.7128**	0.7123	0.5816	**0.5465**	**0.5664**	0.6509	0.5980
	MRPP	**0.6805**	**0.6807**	**0.6586**	0.5901	**0.5653**	**0.5464**	0.6269	0.5592
Textual PopQA	base	0.7127	0.7126	0.6720	0.7293	0.6496	0.6773	0.6106	0.5185
	CoT	**0.6719**	**0.6717**	0.7247	**0.6923**	0.6431	**0.5762**	0.6533	0.6393
	LtM	**0.6077**	**0.6078**	**0.6242**	**0.6052**	**0.5905**	**0.5426**	**0.5771**	0.6089
	MRPP	**0.6908**	**0.6908**	0.7261	**0.6529**	0.6721	**0.5891**	0.6437	0.6048
Semantic PopQA	base	0.6497	0.6456	0.6578	0.5798	0.5916	0.5955	0.6694	0.5964
	CoT	0.7363	0.7359	0.8020	0.7005	0.6449	**0.5541**	0.7513	0.6984
	LtM	**0.6295**	**0.6292**	**0.6467**	**0.5753**	**0.5743**	**0.5381**	**0.5932**	0.6214
	MRPP	0.7295	0.7289	0.7849	0.7035	0.6583	**0.5553**	0.7307	0.6519
Textual TriviaQA	base	0.7225	0.7245	0.7337	0.6822	0.7610	0.7329	0.6212	0.6005
	CoT	**0.7041**	**0.7042**	**0.7039**	0.7283	**0.7090**	**0.6662**	**0.5961**	**0.5830**
	LtM	**0.7024**	**0.7027**	**0.7085**	0.7013	**0.6804**	**0.6714**	**0.6183**	0.6993
	MRPP	**0.7026**	**0.7026**	**0.6880**	0.7325	**0.7436**	**0.6836**	**0.6073**	0.6591
Semantic TriviaQA	base	0.7305	0.7305	0.7424	0.5611	0.6239	0.5866	0.5993	0.6542
	CoT	0.7551	0.7551	**0.7345**	**0.5200**	**0.5836**	**0.5055**	0.7072	0.6766
	LtM	0.7647	0.7650	**0.7259**	0.5902	0.6412	**0.5582**	**0.5960**	**0.6465**
	MRPP	0.7394	0.7395	**0.6958**	0.5750	**0.5927**	**0.5079**	0.6574	0.5804

Table 11: AUROC results of CNN/Daily Mail dataset across different LLMs. Bolded values indicate cases with reduced performance.

		PPL	Sharpness	Eigen Score	SelfCk -Prompt	SelfCk -NLI	Hidden Score	Verbalized Certainty	Attn Score
					Llama-3.1-8B-Instruct				
Textual	base	0.5891	0.5861	0.6295	0.5351	0.5877	0.5139	0.5398	0.6332
	CoT	**0.5789**	**0.5791**	**0.6059**	**0.5001**	**0.5701**	0.5386	**0.5310**	**0.6049**
	LtM	0.6267	0.6240	0.6444	**0.5188**	**0.5376**	0.5425	**0.5354**	**0.6138**
	MRPP	0.6049	0.5982	0.6228	**0.5033**	**0.5517**	0.5335	0.5409	**0.6040**
Semantic	base	0.5681	0.5704	0.6395	0.5260	0.5610	0.5199	0.5287	0.6695
	CoT	**0.5531**	**0.5510**	**0.5957**	0.5278	**0.5258**	**0.5144**	0.5320	**0.6625**
	LtM	0.5873	0.5777	0.6421	0.5521	**0.5172**	0.5374	**0.5085**	**0.6677**
	MRPP	**0.5559**	**0.5527**	**0.5858**	0.5218	**0.5089**	0.5576	**0.5245**	**0.6499**
					Mistral-7B-Instruct-v0.3				
Textual	base	0.6219	0.6219	0.5960	0.5189	0.5314	0.6053	0.5487	0.6404
	CoT	**0.5556**	**0.5556**	**0.5264**	**0.5081**	**0.5150**	**0.5404**	**0.5362**	**0.6033**
	LtM	**0.5785**	**0.5785**	**0.5583**	0.5221	0.5294	**0.5392**	**0.5344**	**0.6061**
	MRPP	**0.5564**	**0.5565**	**0.5216**	**0.5131**	**0.5303**	**0.5244**	**0.5410**	**0.6043**
Semantic	base	0.6132	0.6132	0.5899	0.5256	0.5254	0.5249	0.5316	0.7037
	CoT	**0.5560**	**0.5560**	**0.5296**	**0.5240**	0.5708	**0.5058**	0.5420	0.7155
	LtM	**0.5559**	**0.5559**	**0.5527**	0.5470	0.5458	**0.5141**	**0.5312**	**0.6780**
	MRPP	**0.5315**	**0.5315**	**0.5409**	**0.5202**	0.5403	**0.5243**	**0.5302**	0.7055
					Llama-3.1-70B-Instruct				
Textual	base	0.5554	0.5569	0.5120	0.5612	0.5396	0.5157	0.5313	0.5569
	CoT	**0.5203**	**0.5203**	0.5446	**0.5537**	0.5473	0.5323	**0.5249**	**0.5221**
	LtM	**0.5076**	**0.5074**	0.5142	**0.5451**	**0.5045**	0.5978	**0.5299**	**0.5221**
	MRPP	**0.5107**	**0.5111**	0.5771	0.5891	0.5590	0.5769	0.5421	**0.5120**
Semantic	base	0.5588	0.5583	0.5455	0.5125	0.5261	0.6004	0.5308	0.7052
	CoT	**0.5293**	**0.5289**	**0.5072**	0.5177	0.5298	**0.5893**	**0.5238**	**0.6716**
	LtM	**0.5216**	**0.5219**	**0.5062**	**0.5056**	**0.5248**	**0.5674**	**0.5108**	**0.6806**
	MRPP	**0.5558**	**0.5559**	0.5563	0.5642	0.5649	**0.5368**	**0.5240**	**0.6449**

Table 12: Spearman correlation results of `Llama-3.1-8B-Instruct` across different datasets. The bolded values indicate cases with reduced correlation.

		PPL	Sharpness	Eigen Score	SelfCk -Prompt	SelfCk -NLI	Verbalized Certainty	Hidden Score	Attn Score
Textual	base	-25.10	-24.94	-17.11	31.15	-19.64	46.46	4.89	13.77
	CoT	-29.85	-29.07	-23.91	**19.91**	-31.93	**25.68**	7.52	**6.32**
	LtM	-29.72	-29.42	-21.81	**21.33**	-29.75	**26.53**	**2.71**	**10.20**
	MRPP	-29.07	-28.64	-27.82	**21.32**	-29.53	**24.15**	12.21	**10.74**
Semantic	base	-19.59	-19.27	-28.14	14.93	-13.04	30.52	28.48	18.60
	CoT	-41.67	-41.99	-29.48	**7.53**	**-10.86**	**29.39**	**27.54**	**6.48**
	LtM	-21.30	-21.53	**-23.95**	**8.13**	-17.43	**21.99**	**13.93**	**15.66**
	MRPP	-45.41	-45.58	-35.17	17.97	-17.22	33.96	**27.44**	**10.26**
Textual	base	-40.36	-40.41	-33.04	24.28	-20.54	49.19	26.20	12.07
	CoT	**-37.77**	**-37.27**	-38.04	**7.20**	-32.07	**28.92**	**22.20**	17.55
	LtM	**-35.93**	**-35.64**	-34.32	**14.15**	-32.01	**31.51**	**18.95**	24.74
	MRPP	**-30.51**	**-30.18**	**-32.44**	**16.57**	-30.23	**24.51**	**22.02**	18.39
Semantic	base	-45.00	-45.28	-45.40	4.89	-7.46	62.95	41.78	37.74
	CoT	-47.24	-47.26	-47.26	5.45	-24.11	**40.39**	**36.40**	**24.69**
	LtM	**-44.73**	**-44.92**	**-43.14**	12.11	-41.46	**39.02**	**26.11**	**33.81**
	MRPP	-54.69	-54.79	-56.11	17.77	-43.34	**37.45**	**41.24**	**32.82**
Textual	base	-35.80	-35.44	-28.56	38.10	-45.12	40.24	13.21	7.18
	CoT	**-26.66**	**-26.22**	**-23.66**	**25.94**	**-37.41**	**24.10**	**9.71**	9.20
	LtM	**-28.34**	**-28.18**	**-25.03**	**23.05**	**-29.93**	**24.13**	**12.14**	15.92
	MRPP	**-21.52**	**-21.31**	**-24.11**	**22.52**	**-25.89**	**23.30**	**12.62**	15.38
Semantic	base	-57.87	-57.87	-47.63	17.37	-28.49	32.81	31.09	27.16
	CoT	**-51.38**	**-51.71**	**-35.49**	**5.65**	**-17.42**	**19.28**	**30.22**	**19.60**
	LtM	**-53.37**	**-53.21**	**-40.77**	**13.00**	**-26.63**	**21.69**	**23.98**	**23.72**
	MRPP	-59.79	-59.28	-48.05	29.86	**-26.30**	**27.13**	31.42	**19.20**

Table 13: Spearman correlation results of `Mistral-7B-Instruct-v0.3` across different datasets. The bolded values indicate cases with reduced correlation.

		PPL	Sharpness	Eigen Score	SelfCk -Prompt	SelfCk -NLI	Verbalized Certainty	Hidden Score	Attn Score
Textual HaluEval	base	-27.09	-27.08	-22.92	45.51	-29.37	16.40	11.06	15.26
	CoT	**-24.96**	**-24.97**	-23.76	**34.16**	**-21.90**	17.88	**9.84**	17.62
	LtM	**-23.03**	**-23.04**	-24.26	**30.79**	**-21.87**	19.41	**10.35**	16.08
	MRPP	**-23.08**	**-23.08**	-24.34	**31.10**	**-15.07**	21.79	15.42	18.54
Semantic HaluEval	base	-21.87	-21.86	-22.16	16.47	-16.01	26.81	15.33	20.13
	CoT	-37.38	-37.38	-30.76	16.77	**-9.00**	38.16	16.26	**1.65**
	LtM	-55.59	-55.59	-53.57	34.42	-24.33	45.13	28.04	**18.37**
	MRPP	-38.08	-38.08	-36.07	24.07	**-11.02**	47.15	**13.79**	**1.55**
Textual PopQA	base	-29.69	-29.70	-21.02	44.93	-15.79	29.29	8.10	12.60
	CoT	**-20.61**	**-20.62**	-26.45	**36.14**	**-14.54**	**23.23**	18.63	22.17
	LtM	**-26.11**	**-26.11**	-28.28	**35.18**	-22.19	**23.39**	21.86	24.98
	MRPP	**-17.05**	**-17.05**	-28.91	**34.28**	**-8.12**	**28.59**	24.05	26.37
Semantic PopQA	base	-16.99	-16.99	-13.97	29.24	-22.63	15.92	13.69	19.34
	CoT	-28.93	-28.93	-30.22	**28.55**	-23.69	16.20	16.98	**17.15**
	LtM	-51.36	-51.37	-53.90	39.29	-41.04	27.23	40.45	47.09
	MRPP	-36.38	-36.39	-40.55	34.64	-27.78	32.72	20.07	20.00
Textual TriviaQA	base	-19.28	-19.33	-14.11	46.94	-31.94	20.66	7.20	10.55
	CoT	-23.90	-24.24	-25.62	**38.01**	**-26.62**	**20.61**	15.18	17.40
	LtM	**-18.26**	**-18.40**	-25.12	**35.18**	**-26.91**	23.42	18.75	19.55
	MRPP	-24.87	-24.66	-29.05	**39.61**	**-22.17**	26.29	27.36	26.99
Semantic TriviaQA	base	-27.44	-27.77	-17.31	13.55	-19.10	17.32	11.50	17.15
	CoT	-34.17	-34.25	-36.79	19.73	-29.70	18.67	34.88	**16.45**
	LtM	-50.34	-50.82	-56.95	30.10	-43.43	24.65	44.21	43.06
	MRPP	-51.60	-51.69	-49.91	34.31	-42.28	30.66	40.78	29.99

Table 14: Spearman correlation results of `DeepSeek-R1-Distill-Llama-8B` across different datasets. The bolded values indicate cases with reduced correlation.

		PPL	Sharpness	Eigen Score	SelfCk -Prompt	SelfCk -NLI	Verbalized Certainty	Hidden Score
Textual HaluEval	base	-0.0378	-0.0369	-0.1444	0.1928	-0.2031	0.1440	0.1772
	CoT	**-0.0224**	**-0.0253**	**-0.0302**	0.2285	**-0.1468**	**0.0775**	0.1821
	LtM	**-0.0099**	**-0.0139**	**-0.0315**	0.1994	**-0.1398**	**0.0183**	0.2347
	MRPP	**-0.0212**	**-0.0203**	**-0.0430**	0.2062	**-0.1213**	**0.0530**	0.1975
Semantic HaluEval	base	-0.3112	-0.3104	-0.2987	0.0689	-0.0066	0.0434	0.3913
	CoT	**-0.3107**	**-0.3040**	**-0.1378**	**0.0165**	-0.0120	0.0884	**0.0135**
	LtM	-0.3373	-0.3267	**-0.1399**	**0.0554**	-0.0557	0.0447	**0.0510**
	MRPP	**-0.2884**	**-0.2892**	**-0.1580**	**0.0308**	**-0.0011**	0.0686	**0.0354**
Textual PopQA	base	-0.0377	-0.0366	-0.0492	0.2638	-0.1694	0.0171	0.0249
	CoT	**-0.0137**	**-0.0094**	-0.0535	**0.2518**	**-0.1256**	**0.0555**	**0.2126**
	LtM	**-0.0062**	**-0.0069**	-0.0685	**0.1181**	**-0.0411**	**0.0880**	**0.0628**
	MRPP	**-0.0020**	**-0.0075**	**-0.0393**	**0.0892**	**-0.0331**	**0.0686**	**0.0545**
Semantic PopQA	base	-0.2394	-0.2408	-0.1381	0.1137	-0.0389	0.1260	0.0708
	CoT	-0.3860	-0.3809	-0.3123	**0.0139**	-0.2666	0.3350	**0.0633**
	LtM	-0.2425	**-0.2308**	-0.1677	**0.0060**	-0.1553	0.1619	0.0818
	MRPP	**-0.2014**	**-0.1932**	**-0.1238**	**0.0183**	-0.1239	0.1484	0.1111
Textual TriviaQA	base	-0.0722	-0.0714	-0.1241	0.4061	-0.3693	0.0619	0.0546
	CoT	**-0.0238**	**-0.0209**	**-0.0430**	**0.2098**	**-0.0692**	**0.0605**	0.2513
	LtM	**-0.0301**	**-0.0139**	**-0.0300**	**0.1909**	**-0.0952**	**0.0255**	0.2834
	MRPP	**-0.0410**	**-0.0365**	**-0.0251**	**0.2121**	**-0.0724**	**0.0281**	0.2672
Semantic TriviaQA	base	-0.0237	-0.0231	-0.0419	0.0866	-0.2094	0.2345	0.0126
	CoT	-0.3775	-0.3700	-0.2477	**0.0498**	**-0.1374**	**0.1874**	0.0859
	LtM	-0.4540	-0.4509	-0.3530	**0.0348**	-0.2267	0.2769	0.1660
	MRPP	-0.4320	-0.4251	-0.3374	**0.0134**	-0.2362	0.2881	0.1743
Textual CNN/Daily Mail	base	-0.3460	-0.3437	-0.0701	0.0140	-0.0183	0.2672	0.3268
	CoT	**-0.2458**	**-0.2359**	-0.1568	**0.0055**	-0.0327	**0.0781**	**0.2400**
	LtM	**-0.2476**	**-0.2323**	-0.1624	**0.0011**	-0.0478	**0.1002**	**0.2848**
	MRPP	**-0.2552**	**-0.2445**	-0.1238	**0.0119**	-0.0304	**0.1083**	**0.2755**
Semantic CNN/Daily Mail	base	-0.3661	-0.3631	-0.1145	0.0553	-0.0312	0.0001	0.3559
	CoT	**-0.1898**	**-0.1693**	-0.1297	**0.0057**	-0.0643	0.0427	0.3995
	LtM	**-0.1963**	**-0.1744**	**-0.1141**	**0.0135**	-0.1088	0.0483	0.3868
	MRPP	**-0.1982**	**-0.1768**	-0.1362	**0.0121**	-0.0649	0.0428	0.4083

Table 15: Spearman correlation results of `Llama-3.1-70B-Instruct` across different datasets. The bolded values indicate cases with reduced correlation.

		PPL	Sharpness	Eigen Score	SelfCk -Prompt	SelfCk -NLI	Verbalized Certainty	Hidden Score	Attn Score
Textual HaluEval	base	-0.4222	-0.4222	-0.4078	0.3499	-0.3632	0.2789	0.2739	0.0287
	CoT	**-0.3453**	**-0.3435**	-0.4282	**0.3363**	**-0.1636**	**0.1409**	0.2894	**0.2698**
	LtM	**-0.3669**	**-0.3673**	**-0.3753**	**0.2490**	**-0.1571**	**0.1589**	**0.2312**	**0.2532**
	MRPP	**-0.3914**	**-0.3915**	**-0.4015**	**0.2812**	**-0.2027**	**0.1650**	**0.1693**	**0.2208**
Semantic HaluEval	base	-0.4651	-0.4638	-0.4184	0.1525	-0.2041	0.2288	0.1875	0.1358
	CoT	**-0.4201**	**-0.4201**	**-0.3420**	0.2450	**-0.1057**	**0.0682**	0.3507	0.1896
	LtM	-0.4958	-0.4955	-0.4731	0.1732	**-0.1472**	**0.1527**	0.2608	0.2119
	MRPP	**-0.4128**	**-0.4130**	**-0.3585**	0.1997	**-0.1415**	**0.1019**	0.2616	**0.1297**
Textual PopQA	base	-0.3591	-0.3591	-0.2905	0.2583	-0.2526	0.3334	0.1868	0.3130
	CoT	**-0.2914**	**-0.2912**	-0.3811	0.3261	**-0.2426**	**0.1416**	0.2599	**0.2361**
	LtM	**-0.1838**	**-0.1840**	**-0.2121**	**0.1796**	**-0.1544**	**0.0831**	**0.1316**	**0.1859**
	MRPP	**-0.3251**	**-0.3250**	-0.3852	0.3906	-0.2931	**0.1672**	0.2448	**0.1786**
Semantic PopQA	base	-0.3100	-0.3014	-0.3241	0.1378	-0.2146	0.1757	0.3143	0.2033
	CoT	-0.4383	-0.4375	-0.5465	0.3819	-0.2936	**0.1232**	0.4460	0.3501
	LtM	**-0.2768**	**-0.2764**	**-0.2985**	0.1543	**-0.1794**	**0.0887**	**0.1732**	0.2345
	MRPP	-0.4190	-0.4180	-0.5272	0.3701	-0.2960	**0.0957**	0.4043	0.2700
Textual TriviaQA	base	-0.2103	-0.2121	-0.2208	0.1721	-0.2467	0.2355	0.1145	0.0950
	CoT	**-0.1877**	**-0.1879**	**-0.1876**	**0.2100**	**-0.1922**	**0.1698**	**0.0884**	**0.0763**
	LtM	**-0.1930**	**-0.1933**	**-0.1989**	**0.1920**	**-0.1721**	**0.1910**	**0.1129**	0.1901
	MRPP	**-0.1882**	**-0.1882**	**-0.1746**	**0.2159**	**-0.2262**	**0.1983**	**0.0997**	0.1478
Semantic TriviaQA	base	-0.3406	-0.3407	-0.4543	0.0085	-0.1633	0.1357	0.1771	0.2031
	CoT	-0.5110	-0.5108	-0.4573	0.1115	-0.1818	**0.0083**	0.3921	0.3450
	LtM	-0.5241	-0.5245	**-0.3985**	0.1840	-0.3022	0.1412	0.1937	0.2858
	MRPP	-0.4831	-0.4831	**-0.3878**	0.1527	-0.2058	**0.0456**	0.3112	**0.1851**

Table 16: Spearman correlation results of CNN/Daily Mail dataset across different LLMs. The bolded values indicate cases with reduced correlation.

		PPL	Sharpness	Eigen Score	SelfCk -Prompt	SelfCk -NLI	Hidden Score	Attn Score
				Llama-3.1-8B-Instruct				
Textual	base	-0.1595	-0.1591	-0.2852	-0.0788	-0.1843	-0.0274	0.2857
	CoT	-0.2116	-0.2067	**-0.2452**	**-0.0039**	**-0.1306**	-0.0777	**0.2366**
	LtM	-0.2725	-0.2664	-0.2993	**-0.0438**	**-0.0853**	-0.0885	**0.2423**
	MRPP	-0.2407	-0.2307	-0.2875	**-0.0039**	**-0.1113**	-0.0960	**0.2392**
Semantic	base	-0.1582	-0.1620	-0.2949	-0.0687	-0.1287	-0.0897	0.3541
	CoT	**-0.1192**	**-0.1136**	**-0.2083**	**-0.0488**	**-0.0448**	**-0.0454**	**0.3494**
	LtM	**-0.1941**	**-0.1761**	**-0.2796**	-0.0880	**-0.0339**	**-0.0348**	**0.3452**
	MRPP	**-0.1242**	**-0.1160**	**-0.1790**	**-0.0276**	**-0.0175**	-0.1009	**0.3150**
				Mistral-7B-Instruct-v0.3				
Textual	base	-0.2385	-0.2385	-0.2061	0.0514	-0.0699	-0.2149	0.2910
	CoT	**-0.1323**	**-0.1323**	**-0.0490**	**0.0032**	**-0.0384**	**-0.0641**	**0.2254**
	LtM	**-0.1821**	**-0.1822**	**-0.1269**	0.0678	**-0.0483**	**-0.0633**	**0.2154**
	MRPP	**-0.1140**	**-0.1141**	**-0.0376**	**0.0370**	**-0.0591**	**-0.0350**	**0.2280**
Semantic	base	-0.2227	-0.2226	-0.1985	0.0545	-0.0499	-0.0371	0.4040
	CoT	**-0.1071**	**-0.1070**	**-0.0424**	**0.0389**	-0.1306	**-0.0056**	0.4221
	LtM	**-0.1111**	**-0.1112**	**-0.1075**	0.0860	-0.0802	**-0.0061**	**0.3723**
	MRPP	**-0.0652**	**-0.0652**	**-0.0797**	**0.0270**	-0.0762	-0.0292	0.4113
				Llama-3.1-70B-Instruct				
Textual	base	-0.1201	-0.1227	-0.0165	0.0837	-0.0609	-0.0335	0.1102
	CoT	**-0.0026**	**-0.0017**	-0.0785	0.1075	**-0.0067**	-0.0504	**0.0221**
	LtM	**-0.0127**	**-0.0133**	**-0.0040**	**0.0686**	**-0.0481**	-0.1914	**0.0905**
	MRPP	**-0.0414**	**-0.0419**	-0.1780	**0.0521**	**-0.0263**	-0.1630	**0.0712**
Semantic	base	-0.1044	-0.1033	-0.0924	0.0153	-0.0030	-0.2155	0.4319
	CoT	**-0.0617**	**-0.0606**	**-0.0324**	0.0429	-0.0079	**-0.1417**	**0.3805**
	LtM	**-0.0053**	**-0.0053**	**-0.0269**	0.0169	-0.0719	**-0.1107**	**0.3714**
	MRPP	**-0.0732**	**-0.0729**	-0.1472	0.0887	-0.0621	**-0.0834**	**0.2947**

Table 17: ECE results of `Llama-3.1-8B-Instruct` across different datasets. The bolded values indicate cases with reduced calibration.

		PPL	Sharpness	Eigen Score	SelfCk -Prompt	SelfCk -NLI	Hidden Score	Attn Score
Textual HaluEval	base	16.89	17.00	24.90	8.40	23.64	25.90	40.09
	CoT	14.98	15.03	20.25	**12.05**	17.37	**28.74**	21.90
	LtM	14.93	15.29	**27.23**	**11.27**	14.35	**27.12**	21.45
	MRPP	14.95	15.31	24.70	**11.10**	15.64	**28.64**	18.63
Semantic HaluEval	base	19.50	19.34	13.37	16.08	21.90	25.24	37.27
	CoT	**66.88**	**66.74**	**24.01**	**58.64**	**55.80**	**33.10**	**45.99**
	LtM	16.50	15.59	**20.38**	**17.08**	19.77	21.68	17.76
	MRPP	**65.31**	**65.12**	**20.52**	**63.78**	**64.02**	**34.32**	**50.81**
Textual PopQA	base	20.81	21.18	12.67	9.69	22.75	40.21	53.10
	CoT	18.09	18.28	**13.90**	**18.11**	19.28	31.91	20.50
	LtM	18.80	18.84	**22.04**	**15.69**	17.46	30.82	16.65
	MRPP	18.86	18.82	**29.71**	**12.37**	18.40	33.39	19.04
Semantic PopQA	base	15.29	15.61	4.98	16.66	27.78	27.39	41.15
	CoT	**46.31**	**45.67**	**21.58**	**44.41**	**43.24**	21.19	34.10
	LtM	15.25	15.06	**14.73**	15.83	18.23	24.49	11.62
	MRPP	**47.60**	**47.57**	**22.24**	**56.90**	**52.17**	24.34	40.55
Textual TriviaQA	base	23.98	24.30	27.40	40.62	22.30	10.65	10.93
	CoT	**27.32**	**27.81**	23.50	**42.41**	**23.45**	**16.46**	**29.36**
	LtM	**29.77**	**29.29**	26.69	40.20	**27.11**	**17.08**	**25.70**
	MRPP	**33.83**	**35.21**	23.63	**46.93**	**32.66**	**13.95**	**28.99**
Semantic TriviaQA	base	11.20	12.07	11.90	24.04	21.31	10.47	20.79
	CoT	9.45	11.74	11.56	21.78	**24.11**	**16.33**	13.94
	LtM	8.64	9.38	11.87	19.29	**23.85**	**16.30**	14.38
	MRPP	**12.08**	**12.23**	**13.44**	23.48	20.60	**13.34**	15.41

Table 18: ECE results of `Mistral-7B-Instruct-v0.3` across different datasets. The bolded values indicate cases with reduced calibration.

		PPL	Sharpness	Eigen Score	SelfCk -Prompt	SelfCk -NLI	Hidden Score	Attn Score
Textual HaluEval	base	12.80	12.75	23.43	19.43	17.78	18.71	23.51
	CoT	**17.06**	**17.10**	**26.82**	**38.75**	**19.96**	**18.93**	14.34
	LtM	**18.26**	**18.24**	**29.60**	**40.78**	**18.42**	**20.62**	14.58
	MRPP	**18.63**	**18.67**	**30.54**	**43.73**	**20.05**	**19.71**	13.19
Semantic HaluEval	base	11.26	11.23	19.60	26.37	23.74	14.24	18.15
	CoT	**56.92**	**56.91**	17.20	22.25	**61.50**	**47.64**	**56.51**
	LtM	**60.89**	**60.89**	**25.75**	21.92	**67.90**	**46.80**	**56.03**
	MRPP	**55.47**	**55.43**	17.86	20.97	**68.08**	**48.73**	**58.77**
Textual PopQA	base	17.76	17.74	26.49	22.52	28.24	21.89	25.30
	CoT	**19.12**	**19.15**	22.59	**40.93**	20.86	19.84	14.76
	LtM	**21.59**	**21.61**	**31.86**	**42.11**	18.28	18.43	11.04
	MRPP	**22.19**	**22.18**	**30.76**	**47.66**	22.37	20.88	12.72
Semantic PopQA	base	14.19	14.15	18.56	22.49	25.23	15.83	18.74
	CoT	**55.14**	**55.15**	17.28	**23.01**	**60.14**	**44.88**	**53.89**
	LtM	**56.82**	**56.83**	**26.16**	22.47	**62.59**	**47.89**	**60.39**
	MRPP	**51.42**	**51.44**	15.97	21.37	**65.40**	**49.18**	**60.06**
Textual TriviaQA	base	31.75	31.82	13.52	6.04	30.41	25.51	22.26
	CoT	29.45	29.37	**19.38**	**12.32**	**32.03**	**30.01**	**32.30**
	LtM	**33.90**	**33.42**	**28.24**	**13.69**	**31.64**	**26.12**	**33.15**
	MRPP	26.24	26.14	**23.55**	**15.48**	**33.93**	21.80	**31.19**
Semantic TriviaQA	base	11.45	11.98	23.09	36.09	25.21	14.18	14.59
	CoT	**21.23**	**21.51**	15.00	27.66	**35.64**	**18.23**	**24.38**
	LtM	**26.65**	**26.63**	**25.76**	16.83	**40.52**	**25.51**	**33.13**
	MRPP	**23.87**	**24.06**	21.50	19.57	**37.18**	**21.20**	**31.04**

Table 19: ECE results of `DeepSeek-R1-Distill-Llama-8B` across different datasets. The bolded values indicate cases with reduced calibration.

		PPL	Sharpness	Eigen Score	SelfCk -Prompt	SelfCk -NLI	Verbalized Certainty	Hidden Score
Textual HaluEval	base	0.2133	0.2098	0.4199	0.5159	0.3615	0.3567	0.4295
	CoT	0.2088	0.2060	**0.4256**	**0.6087**	0.1700	0.3016	**0.4706**
	LtM	0.1774	0.1789	0.4124	**0.6148**	0.1575	0.3186	**0.4798**
	MRPP	**0.2133**	0.2053	0.4124	**0.6052**	0.1911	0.3225	**0.4860**
Semantic HaluEval	base	0.2590	0.2607	0.1069	0.1943	0.2234	0.1299	0.1212
	CoT	**0.4416**	**0.4446**	**0.2364**	0.0880	**0.7213**	**0.2962**	**0.1452**
	LtM	**0.4757**	**0.4789**	**0.2534**	0.0805	**0.7346**	**0.3158**	**0.1396**
	MRPP	**0.4711**	**0.4763**	**0.2551**	0.0817	**0.7330**	**0.2981**	**0.1345**
Textual PopQA	base	0.2064	0.2054	0.4866	0.3780	0.4322	0.3659	0.5111
	CoT	**0.2726**	**0.2647**	0.4598	**0.6850**	0.1289	0.3245	0.4400
	LtM	**0.3204**	**0.3138**	**0.5445**	**0.7279**	0.1579	**0.3846**	0.4601
	MRPP	**0.3286**	**0.3196**	**0.5573**	**0.7268**	0.1879	**0.3744**	0.4461
Semantic PopQA	base	0.2273	0.2270	0.2293	0.2563	0.3350	0.2059	0.1713
	CoT	0.1942	0.2071	0.1469	0.2231	**0.4832**	0.1389	0.1533
	LtM	0.2247	**0.2355**	0.1348	0.2221	**0.5359**	0.1425	0.1501
	MRPP	**0.2445**	**0.2472**	0.1421	0.2138	**0.5690**	0.1619	**0.1721**
Textual TriviaQA	base	0.1205	0.1216	0.1923	0.2668	0.1416	0.1565	0.2142
	CoT	**0.2509**	**0.2548**	**0.2249**	**0.2752**	**0.4003**	**0.1632**	0.1546
	LtM	**0.2644**	**0.2795**	**0.2367**	**0.2925**	**0.3937**	**0.1656**	0.1626
	MRPP	**0.2697**	**0.2745**	**0.2267**	**0.2866**	**0.4178**	**0.1607**	0.1509
Semantic TriviaQA	base	0.2113	0.2074	0.2000	0.2514	0.2351	0.1268	0.1734
	CoT	**0.3014**	**0.3050**	0.1707	0.1635	**0.6069**	**0.2051**	0.1071
	LtM	**0.3645**	**0.3699**	**0.2390**	0.1186	**0.6572**	**0.2456**	0.0904
	MRPP	**0.3463**	**0.3470**	**0.2176**	0.1219	**0.6215**	**0.2232**	0.0784
Textual CNN/Daily Mail	base	0.1048	0.1093	0.3753	0.4203	0.3140	0.1259	0.2378
	CoT	**0.1534**	**0.1503**	0.2742	0.2709	**0.4309**	**0.1850**	0.1279
	LtM	**0.1149**	**0.1144**	0.2721	0.3257	**0.3285**	**0.1443**	0.1390
	MRPP	**0.1232**	**0.1176**	0.3332	0.3311	**0.3695**	**0.1538**	0.1513
Semantic CNN/Daily Mail	base	0.0950	0.0969	0.2595	0.3013	0.4183	0.1774	0.1289
	CoT	**0.1321**	**0.1298**	**0.4058**	**0.3618**	0.3593	**0.1851**	**0.1649**
	LtM	**0.1184**	**0.1227**	**0.4462**	**0.3650**	0.3173	**0.1869**	**0.1987**
	MRPP	**0.1179**	**0.1286**	**0.3661**	**0.4099**	0.3243	**0.1781**	**0.1988**

Table 20: ECE results of `Llama-3.1-70B-Instruct` across different datasets. The bolded values indicate cases with reduced calibration.

		PPL	Sharpness	Eigen Score	SelfCk -Prompt	SelfCk -NLI	Hidden Score	Attn Score
Textual HaluEval	base	0.1653	0.1650	0.1446	0.0923	0.1761	0.3062	0.3536
	CoT	**0.2211**	**0.2218**	**0.3481**	**0.2309**	**0.2143**	**0.4463**	**0.3715**
	LtM	**0.1930**	**0.1918**	**0.2600**	**0.2045**	0.1752	**0.4767**	**0.3642**
	MRPP	**0.2225**	**0.2228**	**0.3641**	**0.2463**	**0.2109**	**0.4350**	**0.4757**
Semantic HaluEval	base	0.1075	0.0977	0.1295	0.2224	0.2170	0.1082	0.2587
	CoT	**0.1507**	**0.1493**	0.1218	0.2018	0.2384	0.1038	0.1468
	LtM	**0.1828**	**0.1989**	**0.1495**	**0.2231**	0.3455	**0.1234**	0.1492
	MRPP	**0.1292**	**0.1329**	0.1269	0.2217	0.2551	**0.1276**	0.1764
Textual PopQA	base	0.1507	0.1508	0.1232	0.0992	0.1803	0.2715	0.4548
	CoT	0.1342	0.1329	**0.1616**	**0.2019**	**0.2204**	0.1043	0.1204
	LtM	**0.1718**	**0.1710**	**0.2060**	**0.1233**	0.1286	**0.2858**	0.2076
	MRPP	0.1417	0.1391	**0.1390**	**0.2024**	**0.1834**	0.1152	0.1195
Semantic PopQA	base	0.1322	0.1274	0.0769	0.1423	0.2190	0.1820	0.4065
	CoT	**0.1677**	**0.1657**	**0.1330**	0.0841	0.2075	**0.2371**	0.1435
	LtM	**0.1389**	**0.1363**	**0.1644**	0.1382	**0.2567**	**0.2177**	0.1499
	MRPP	**0.1756**	**0.1760**	**0.0848**	0.0791	0.2172	**0.2429**	0.1858
Textual TriviaQA	base	0.4652	0.4484	0.4546	0.4893	0.4617	0.2768	0.1262
	CoT	**0.4976**	**0.4997**	0.3119	0.3898	**0.4924**	**0.3038**	**0.3421**
	LtM	0.4427	**0.4564**	0.3899	0.4210	**0.5202**	0.2436	**0.3003**
	MRPP	**0.4802**	**0.4802**	0.3111	0.3881	**0.4868**	0.2489	**0.3546**
Semantic TriviaQA	base	0.1657	0.1853	0.0907	0.2557	0.2150	0.1129	0.1556
	CoT	0.1598	0.1565	**0.1015**	0.1674	**0.2228**	**0.1809**	0.1515
	LtM	0.1060	0.1112	0.0739	0.1364	0.1678	**0.1883**	0.1406
	MRPP	0.1373	0.1384	**0.1181**	0.1707	0.1564	**0.2441**	**0.1867**

Table 21: ECE results of CNN/Daily Mail dataset across different LLMs. The bolded values indicate cases with reduced calibration.

		PPL	Sharpness	Eigen Score	SelfCk -Prompt	SelfCk -NLI	Hidden Score	Attn Score
				Llama-3.1-8B-Instruct				
Textual	base	0.1529	0.1487	0.1974	0.1831	0.1882	0.1557	0.2329
	CoT	0.1209	0.1227	0.1929	**0.2008**	**0.2377**	0.1553	0.1951
	LtM	0.1321	0.1198	0.1484	**0.1952**	**0.2605**	**0.1896**	0.1880
	MRPP	0.1237	0.1319	**0.2409**	**0.1903**	**0.2169**	**0.1658**	0.2272
Semantic	base	0.1805	0.1753	0.1629	0.1897	0.2506	0.1750	0.1284
	CoT	0.1570	0.1568	**0.3774**	0.1668	0.2370	**0.2014**	**0.2173**
	LtM	0.1534	0.1510	0.1433	0.1677	**0.2636**	0.1693	**0.2079**
	MRPP	0.1611	0.1584	**0.2984**	**0.2091**	0.2377	0.1626	**0.2358**
				Mistral-7B-Instruct-v0.3				
Textual	base	0.0995	0.0995	0.1709	0.4466	0.2365	0.1166	0.2240
	CoT	**0.1536**	**0.1536**	**0.2428**	0.3138	**0.3571**	**0.1816**	0.1310
	LtM	**0.1850**	**0.1853**	**0.2722**	0.3012	**0.3653**	**0.1694**	0.1188
	MRPP	**0.1553**	**0.1553**	**0.2503**	0.3414	**0.3042**	**0.1868**	0.1393
Semantic	base	0.1391	0.1396	0.1576	0.3481	0.3139	0.1721	0.1273
	CoT	**0.1646**	**0.1649**	**0.3353**	**0.5001**	0.1607	**0.1898**	0.2619
	LtM	**0.1757**	**0.1750**	**0.4340**	**0.4852**	0.2088	**0.1857**	0.2578
	MRPP	**0.1786**	**0.1780**	**0.3517**	**0.5160**	0.2510	0.1709	0.2737
				Llama-3.1-70B-Instruct				
Textual	base	0.2342	0.2350	0.2560	0.2136	0.2281	0.1680	0.2037
	CoT	0.1495	0.1525	0.2559	**0.2533**	0.2014	**0.2092**	**0.2457**
	LtM	0.1484	0.1494	**0.2571**	**0.2168**	0.1717	0.1676	**0.2456**
	MRPP	0.1775	0.1754	**0.2691**	0.2010	0.2011	0.1471	**0.2139**
Semantic	base	0.2467	0.2452	0.2649	0.2374	0.2475	0.1620	0.1010
	CoT	0.2011	0.1982	0.2588	**0.2665**	0.2052	**0.1829**	**0.1847**
	LtM	0.2005	0.2002	0.2358	**0.2731**	0.1759	**0.1925**	**0.1757**
	MRPP	0.2190	0.2138	**0.3725**	0.2204	0.2160	**0.2248**	**0.1947**

Table 22: AUROC results of TruthfulQA. The bolded values indicate cases with reduced performance.

		PPL	Sharpness	Eigen Score	SelfCk -Prompt	SelfCk -NLI	Verbalized Certainty	Hidden Score	Attn Score
					Llama-3.1-8B-Instruct				
Info.	base	65.37	65.74	62.61	55.57	56.28	61.70	57.74	52.13
	CoT	**54.87**	**54.61**	**54.08**	**50.22**	56.55	**53.64**	**50.82**	57.46
	LtM	**53.69**	**53.49**	66.49	56.37	**55.95**	**57.03**	61.21	60.56
	MRPP	**56.15**	**55.81**	**55.76**	55.62	63.02	70.01	**51.33**	80.12
Truth.	base	50.35	50.36	54.34	59.17	60.31	54.69	52.43	50.78
	CoT	53.06	53.61	**51.67**	**54.79**	**58.23**	**53.23**	54.84	**50.14**
	LtM	53.16	53.63	**53.18**	**56.15**	62.80	**52.35**	57.10	52.53
	MRPP	54.32	54.65	**51.22**	**52.42**	**55.40**	57.67	53.73	**50.72**
					Mistral-7B-Instruct-v0.3				
Info.	base	51.94	51.91	75.43	63.71	66.54	69.46	62.21	70.23
	CoT	76.82	76.82	75.77	**62.05**	70.74	76.79	**52.47**	**55.91**
	LtM	74.63	74.63	78.59	69.66	68.55	79.66	**59.51**	**58.12**
	MRPP	74.23	74.22	**69.51**	67.95	66.66	80.82	**56.80**	**62.11**
Truth.	base	54.27	54.26	57.79	56.59	62.20	58.73	55.92	56.20
	CoT	**53.07**	**53.06**	59.79	59.85	**57.53**	62.39	**51.40**	**54.98**
	LtM	55.42	55.41	63.16	59.98	**56.10**	**56.83**	**50.57**	**53.26**
	MRPP	**50.43**	**50.43**	**57.25**	59.71	**56.19**	**55.64**	**52.88**	**51.13**

Table 23: To mitigate potential biases introduced by the LLaMA-based judge model, we also fine-tune the evaluation model using Qwen. The evaluation results based on the Qwen judge model are consistent with our previous findings, confirming the robustness of our conclusions.

		PPL	Sharpness	Eigen Score	SelfCk -Prompt	SelfCk -NLI	Verbalized Certainty	Hidden Score	Attn Score
					Llama-3.1-8B-Instruct				
Info.	base	56.02	56.52	61.56	54.35	65.94	67.57	72.12	71.36
	CoT	**54.04**	**55.31**	72.77	**53.61**	**61.02**	**66.32**	**51.55**	74.78
	LtM	59.23	58.75	65.92	**50.21**	**57.74**	**59.63**	**64.98**	**58.63**
	MRPP	58.6	62.07	67.38	82.95	**55.03**	**52.00**	**70.88**	**69.74**
Truth.	base	56.31	57.33	67.49	56.06	53.19	52.64	75.18	77.3
	CoT	56.76	**55.54**	**64.28**	**52.61**	**50.32**	**52.31**	**60.81**	**63.94**
	LtM	58.73	58.66	**59.91**	**53.21**	54.24	52.81	**56.31**	**59.4**
	MRPP	57.75	**57.11**	**66.35**	**52.55**	56.59	**52.07**	**64.59**	**68.54**
					Mistral-7B-Instruct-v0.3				
Info.	base	56.03	56.07	53.91	69.34	78.91	74.14	89.07	85.9
	CoT	69.61	69.67	60.44	80.17	**74.66**	**67.95**	**82.08**	**53.69**
	LtM	81.03	80.99	80.95	**69.88**	**63.72**	**61.17**	**52.1**	**64.38**
	MRPP	75.29	75.29	68.44	81.81	**65.99**	**51.71**	**74.69**	**77.45**
Truth.	base	58.76	58.73	52.31	54.12	62.17	62.23	75.86	75.73
	CoT	59.41	59.42	**51.59**	59.68	64.38	**58.91**	**67.59**	**65.51**
	LtM	**58.47**	**58.47**	**50.22**	54.89	**59.5**	66.14	**66.21**	**68.41**
	MRPP	**58.58**	**58.58**	**52.12**	59.02	62.68	**61.36**	**64.21**	**71.65**