Calibrating LLM Confidence with Semantic Steering: A Multi-Prompt Aggregation Framework

Anonymous ACL submission

Abstract

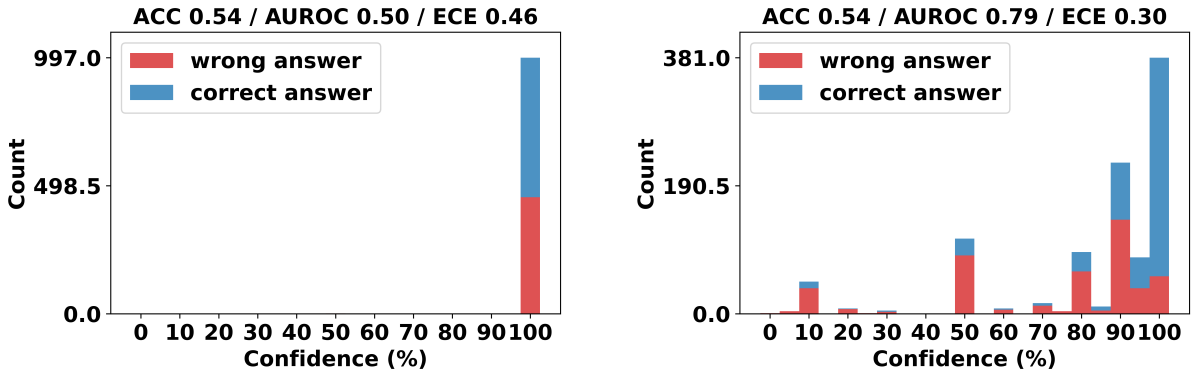
Large Language Models (LLMs) often exhibit misaligned confidence scores, usually overestimating the reliability of their predictions. While verbalized confidence in Large Language Models (LLMs) has gained attention, prior work remains divided on whether confidence scores can be systematically steered through prompting. Recent studies even argue that such prompt-induced confidence shifts are negligible, suggesting LLMs' confidence calibration is rigid to linguistic interventions. Contrary to these claims, we first rigorously confirm the existence of directional confidence shifts by probing three models (including GPT3.5, LLAMA3-70b, GPT4) across 7 benchmarks, demonstrating that explicit instructions can inflate or deflate confidence scores in a regulated manner. Based on this observation, we propose a novel framework containing three components: confidence steering, steered confidence aggregation and steered answers selection, named SteeringConf. Our method, SteeringConf, leverages a confidence manipulation mechanism to steer the confidence scores of LLMs in several desired directions, followed by a summarization module that aggregates the steered confidence scores to produce a final prediction. We evaluate our method on 7 benchmarks and it consistently outperforms the baselines in terms of calibration metrics in task of confidence calibration and failure detection.

1 Introduction

Large Language Models (LLMs) have revolutionized artificial intelligence by achieving remarkable performance across diverse tasks, from text generation to complex reasoning (Brown et al., 2020; Wei et al., 2022; Petroni et al., 2019). However, their practical deployment faces a critical challenge: misaligned confidence calibration (Jiang et al., 2021; Lin et al., 2022; Shrivastava et al., 2023). LLMs often produce overconfident predictions (Xiong et al., 2023; Tian et al., 2023) that do not reflect their true likelihood of being correct, raising concerns about their reliability in high-stakes applications such as healthcare (Bedi et al., 2024; Savage et al., 2024), legal analysis (Guha et al., 2023), and autonomous systems (Chen et al., 2024; Wang et al., 2024). While prior work has explored verbalized confidence—probing LLMs to self-assess their prediction certainty (Xiong et al., 2023; Tian et al., 2023), the field remains divided on two pivotal questions: Can linguistic interventions, such as prompting, systematically steer an LLM's confidence scores in a controlled manner? And if confidence scores can be steered, can we utilize this steering to get a better calibrated confidence?

Recent studies (Xiong et al., 2023) argue that prompt-induced confidence shifts are negligible, positing that LLMs' calibration is inherently rigid and resistant to linguistic steering. This perspective, however, conflicts with our observations as shown in Figure 1. To resolve this contradiction, we conduct a rigorous empirical investigation on seven benchmarks in Section 6. Our experiments systematically confirm that explicit instructions (e.g., "Be *very cautious*" or "Be *very confident*") induce directional confidence shifts, while mild instructions (e.g., "Be *cautious*" or "Be *confident*") can not induce desired confidence shifts. This finding challenges the prevailing assumption of calibration rigidity and opens new avenues for improving LLM trustworthiness.

Building on this insight, we propose SteeringConf, a novel framework for dynamic confidence calibration. SteeringConf comprises three components: (1) a confidence steering mechanism that steers LLM confidence in specified directions (e.g., conservative or optimistic calibration) through tailored prompts, (2) an aggregation module that aggregates multiple steered confidences to produce a final, better-calibrated output based on the consistency of multiple steered answers and associated confidences, (3) and an selection criteria to



(a) Vanilla Verbalized Confidence

(b) Verbalized Confidence Steered to be *very cautious*

Figure 1: The comparison between vanilla verbalized confidence and *very cautious* Prompt Steered Confidence on Object Counting Dataset with GPT-3.5 as LLM. One can see that vanilla verbalized confidence has an extreme overconfidence issue, while our *very cautious* prompt successfully steered confidence to a better calibrated distribution. Moreover, the calibration performance of steered confidence method is much better than vanilla version: 29% improvement over AUROC and 16% improvement over ECE.

choose the steered answer associated with best confidence in align with previous calibrated confidence. Evaluated across seven benchmarks spanning professional knowledge, common-sense, ethics, and reasoning tasks and combined with three state-of-the-art models (GPT-3.5, LLaMA 3, GPT-4) , SteeringConf consistently outperforms existing calibration methods in both confidence calibration, e.g., reducing Expected Calibration Error (ECE) by up to 39.8%, and failure detection, e.g., improving AUROC by 33.9%.

2 Related Work

Confidence from External Knowledge This paradigm leverages external knowledge through model-agnostic approaches: Proxy models employ lightweight neural networks trained on synthetic Q/A/confidence datasets (Tsai et al., 2024) or model internal states (Mielke et al., 2022), though constrained by training data limitations. Human feedback mechanisms demonstrate reliability through self-repair systems (Giulianelli et al., 2023) but face scalability challenges. Knowledge tool integration combines search engines and code interpreters (Gou et al., 2023; Chern et al., 2023) at a significant computational cost.

Confidence from Logits This method exploits model-specific internal computations. Logit-based methods aggregate token probabilities either through full-sequence likelihood (Jiang et al., 2021; Si et al., 2022) or answer-specific token selection (Ye et al., 2024), fundamentally limited by semantic disconnection between token probabilities and high-level uncertainty (Kuhn et al., 2022; Wang and Holmes, 2024; Lin et al., 2022). And the alignment procedure (OpenAI, 2024) could also ruin the quality of logits for calibration (Tian et al., 2023).

Verbalized Confidence This approach directly queries the LLM to self-assess and articulate its confidence through natural language expressions. This method is model-agnostic, requiring only black-box access to the LLM while maintaining low computational overhead (constant token expansion). Current implementations primarily measure confidence in factual correctness (Tian et al., 2023; Chen and Mueller, 2023), explanation confidence (Kadavath et al., 2022; Tanneru et al., 2023). Recent explorations also investigate calibrating linguistic uncertainty markers (*e.g.*, "probably" vs "certainly") (Mielke et al., 2022; Zhou et al., 2023). (Xiong et al., 2023) summarize a unified framework considering sample consistency, which are most close to our method.

3 Preliminary

3.1 Large Language Models with Prompting

We formally define a large language model (LLM) as a generative function $M : \mathcal{X} \to \mathcal{X}$, where $\mathcal{X}$ denotes the text space encompassing all possible textual inputs and outputs. The prompting mechanism $P : \mathcal{X} \to \mathcal{X}$ operates as a template transformation function that maps an original input $x \in \mathcal{X}$ to an instruction-augmented prompt. This process enables explicit guidance of model behavior through

carefully engineered input formulations.

One general generation pipeline of LLM to tackle the task comprises three key components:

1. *Prompt Engineering*: Application of P to create task-specific input prompts.

2. *Text Generation*: Execution of large language model M to produce textual outputs.

3. *Information Extraction*: Implementation of readout function $R : \mathcal{X} \to \mathcal{Y}$ that maps generated text to structured predictions in output space $\mathcal{Y}$.

The composite prediction function can be expressed as:

$$f = R \circ M \circ P : \mathcal{X} \to \mathcal{Y}, \tag{1}$$

where $\circ$ denotes function composition. Notably, readout functions typically employ rule-based methods (e.g., regular expressions) and demonstrate task-specific dependence on prompting strategies.

3.2 Verbalized Confidence

Confidences are quantitative measurement of the reliability and uncertainty of predictions made by LLMs. If the confidence score is close to 1, it indicates that the model is confident in its prediction, while a confidence score close to 0 indicates that the model is uncertain about its prediction. There are various ways to compute confidence scores; we focus on the verbalized confidence score by the LLMs (Xiong et al., 2023; Tian et al., 2023). To elicit the confidence score from the LLM, we first change the prompting template function P to P_{conf} that asks the LLM to output the prediction and the confidence score. Then we change the readout function R to R_{conf} that extracts output text of LLM M and returns two values: the prediction and the confidence score. The prediction and the confidence score can be denoted as $\big(f(x), c(x)\big) = R_{\text{conf}} \circ M \circ P_{\text{conf}}(x)$, where $c(x)$ is the confidence score of the prediction $f(x)$.

4 Method

Building upon the unified framework for verbalized confidence elicitation proposed in (Xiong et al., 2023), which comprises Prompting, Sampling and Aggregation phases, we present SteeringConf - a calibrated confidence estimation framework through systematic prompt steering. Our key insight stems from the observation that LLM confidence scores exhibit directional sensitivity to semantic perturbations in prompting. This motivates our three-stage approach: 1) confidence steering through semantic prompt variations that combines the Prompting and Sampling phases. 2) aggregation of steered confidence. 3) selection of steered answer. This tribal-phase architecture enables both fine-grained confidence adjustment and robust uncertainty quantification. The overview of our method is summarized in Figure 2.

4.1 Confidence Steering

The steering mechanism begins with constructing a symmetric prompt set. We first define a set of K steering prompts $\{P_{\text{conf}}^{-K}, P_{\text{conf}}^{-K+1}, \ldots, P_{\text{conf}}^{0}, \ldots, P_{\text{conf}}^{K-1}, P_{\text{conf}}^{K}\}$, where steering magnitude $K \in \mathbb{Z}^+$ controls the perturbation intensity. Each prompt P_{conf}^{k} embeds distinct confidence directives from a specified spectrum. In our implementation, we apply a simple but effective steering magnitude setting $K = 2$ as a moderate granularity, denoted as $\{$*very cautious*, *cautious*, *vanilla*, *confident*, *very confident* $\}$.

Given an input text x, we apply the steering prompt set $\{P_{\text{conf}}^{k}\}_{k=-K}^{K}$ to the LLM, generating $2K + 1$ prediction-confidence pairs. Formally, we obtain $\{f_k(x), c_k(x)\}_{k=-K}^{K}$ which is defined as

$$(f_k(x), c_k(x)) \triangleq R_{\text{conf}} \circ M \circ P_{\text{conf}}^{k}(x)$$

Note: While we hypothesize directional monotonicity ($c_{-K}(x) < \cdots < c_0(x) < \cdots < c_K(x)$), real-world observations may deviate from strict monotonicity due to LLMs' complex response patterns. Our aggregation module therefore requires robustness to such non-ideal cases.

4.2 Steered Confidence Aggregation

Given the steering-induced predictions $\{f_k(x)\}_{k=-K}^{K}$ and associate confidence scores $\{c_k(x)\}_{k=-K}^{K}$, our aggregation framework synthesizes three complementary signals to produce calibrated confidence estimates. The design philosophy stems from two key observations: (1) prediction consistency across steering directions reflects model certainty, and (2) confidence consistency under steering indicates calibration reliability. We formalize this through three synergistic components:

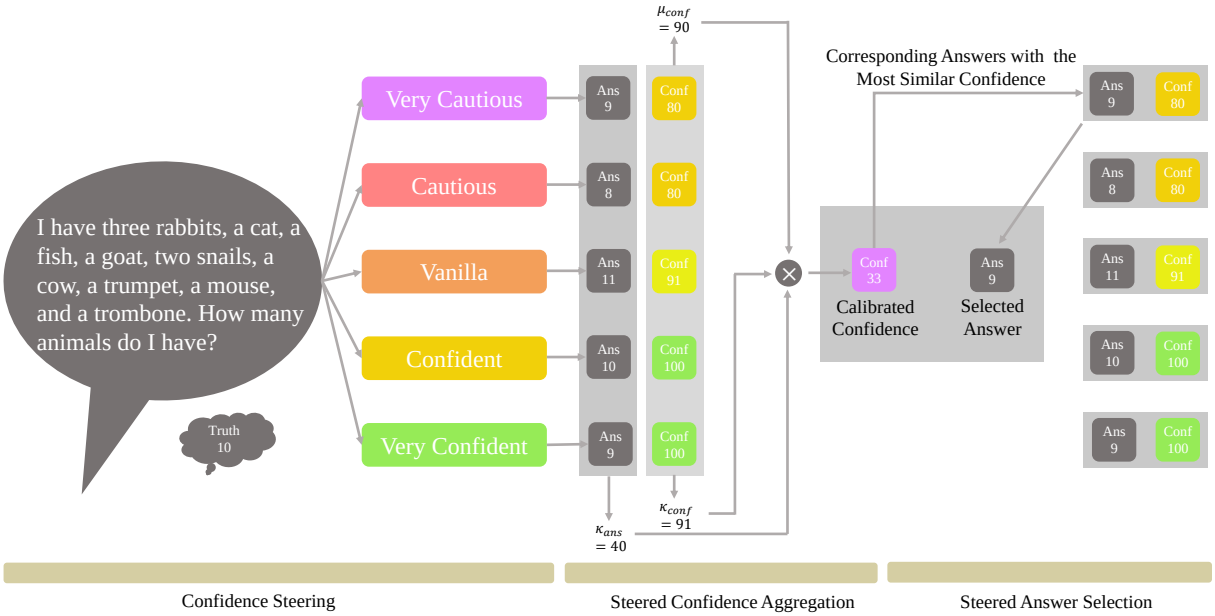


Figure 2: An overview and an instantiation of our SteerConf framework, which consists of three components: Confidence Steering, Steered Confidence Aggregation and Steered Answer Selection. By using a spectrum of $2K+1$ steering prompts: {Very Cautious, Cautious, Vanilla, Confident, Very Confident}, we firstly process the question with these $2K+1$ steering prompts. Then we obtain $2K+1$ pairs of answer and confidence. Next, we aggregate the answers and confidences into one calibrated confidence by their consistency. Finally, we select one steered answer with its corresponding confidence that is closest to the calibrated confidence.

1. **Answer Consistency**: When LLMs produce divergent predictions under semantic perturbations, this signal predict its inherent uncertainty. We quantify answer consistency through prediction frequency as in (Xiong et al., 2023):

$$\text{freq}(y) = \frac{1}{2K+1} \sum_{k=-K}^{K} \mathbb{I}(f_k(x) = y) \quad (2)$$

The dominant prediction $f_m(x) = \arg\max_y \text{freq}(y)$ and its consistency score $\kappa_{\text{ans}} = \text{freq}(f_m(x))$ define the first calibration factor. Higher κ_{ans} indicates stronger agreement and stability across steering directions, while lower κ_{ans} values indicate conflicting predictions across steering directions, suggesting inherent model uncertainty.

2. **Confidence Consistency**: While answer consistency evaluates prediction stability, we introduce confidence consistency to assess the reliability of confidence scores under steering perturbations. This novel component analyzes both central tendency (Mean) and dispersion (Std Dev) :

$$\mu_c = \frac{1}{2K+1} \sum_{k=-K}^{K} c_k(x) \quad (3)$$

$$\sigma_c = \sqrt{\frac{1}{2K+1} \sum_{k=-K}^{K} (c_k(x) - \mu_c)^2} \quad (4)$$

The confidence consensus score then combines these statistics:

$$\kappa_{\text{conf}} = \frac{1}{1 + \sigma_c/\mu_c} \quad (5)$$

This formulation serves three purposes: (1) bounded output in $(0, 1]$ as a scaling factor, (2) Penalizes high variance (σ_c) relative to mean confidence (μ_c), and (3) direct scaling with mean confidence μ_c. High κ_{conf} values indicate steering-invariant confidence estimates, suggesting well-calibrated certainty.

3. **Calibrated Confidence**: The final confidence estimate combines our consensus metrics through multiplicative interaction:

$$c(x) = \mu_c \cdot \kappa_{\text{ans}} \cdot \kappa_{\text{conf}} \quad (6)$$

This formulation naturally downweights the raw confidence average μ_c when either answer

inconsistency ($\kappa_{\text{ans}} \downarrow$) or confidence instability ($\kappa_{\text{conf}} \downarrow$) occurs. And it will preserve well-calibrated confidence estimate μ_c when both consistency metrics are high, indicating that raw confidence average is reliable.

4.3 Steered Answer Selection

To harmonize confidence estimates with prediction choices, we develop a prediction selection mechanism in this section. The intuition is that steering directions with confidence values closest to our calibrated confidence $c(x)$ should dominate the final prediction. We first map $c(x)$ to the steering index space through linear quantization:

$$j = \left\lfloor \frac{c(x) - c_{\min}}{c_{\max} - c_{\min}} \cdot n_{\text{bins}} \right\rfloor \tag{7}$$

$$f(x) = \begin{cases} f_{-K}(x) & \text{if } j < -K, \\ f_j(x) & \text{if otherwise,} \end{cases} \tag{8}$$

where $c_{\min} = \min_k c_k(x)$, $c_{\max} = \max_k c_k(x)$. Since $c(x) \leq \mu_c \leq c_{max}$, we have $j \leq K$, therefore we only condition corner case of $j < -K$.

To summarize, this complete aggregation process achieves dual calibration: confidence estimates are refined through consistency analysis, while predictions are selected through confidence-aware steering alignment. It's noted that the confidence consistency component is particularly crucial for handling LLMs' tendency toward overconfidence - high variance in confidences under steering automatically triggers attenuation through κ_{conf}. The prediction selection part bridges the gap between continuous confidence calibration and discrete prediction choices.

5 Main Experiment

5.1 Setup

Datasets We assess confidence estimation quality across five reasoning categories: (1) *Commonsense Reasoning* using Sports Understanding (SportUND) (Kim, 2021) and StrategyQA (Geva et al., 2021) from BigBench (Ghazal et al., 2013); (2) *Arithmetic Reasoning* evaluated on GSM8K (Cobbe et al., 2021); (3) *Symbolic Reasoning* covering Date Understanding (DateUnd) (Wu and Wang, 2021) and Object Counting (Object-Cou) (Wang et al., 2019) (BigBench); (4) *Professional Knowledge* tested through Law (Prf-Law) from MMLU (Hendrycks et al., 2021); and (5) *Ethical Knowledge* examined via Business Ethics (Biz-Ethics) in MMLU (Hendrycks et al., 2021).

Models We use several widely-utilized large language models (LLMs) of varying sizes, such as GPT-3.5 (OpenAI, 2021), Llama 3 (AI@Meta, 2024), and GPT-4 (OpenAI, 2024). Specifically, Llama 3 employs a 70B size with 4-bit quantization, while GPT-3.5 and GPT-4 were accessed between August 1, 2024, and February 1, 2025. We compare our method with vanilla verbalized confidence elicited from these models in Table 1.

Baselines We also compare sampling and prompting based methods summarized as in (Xiong et al., 2023) including Misleading, Self-Random, Prompt and Topk from (Tian et al., 2023) in Table 2. As the same setting in their original paper (Xiong et al., 2023), we adopt GPT-3.5 as the LLM backbone, and they all use CoT, while consistency aggregation metric (Xiong et al., 2023) is used for these baselines.

5.2 Metrics

To assess the quality of confidence estimates produced by models, two distinct but complementary evaluation frameworks are commonly used: calibration analysis and failure prediction (Xiong et al., 2023). Calibration examines the alignment between a model's stated confidence levels and its empirical accuracy—for instance, predictions made with 80% confidence should ideally exhibit an 80% accuracy rate. These well-calibrated estimates are particularly important in contexts such as risk evaluation. In contrast, failure prediction tasks evaluate a model's ability to rank confidence scores such that correct predictions receive higher values than incorrect ones, testing whether confidence metrics can reliably separate accurate from inaccurate outputs. For this work, Expected Calibration Error (ECE) serves as the primary calibration metric, while failure prediction performance is measured using the Area Under the Receiver Operating Characteristic Curve (AUROC). To address potential class imbalance arising from differences in accuracy across samples, we additionally incorporate AUPRC-Positive (PR-P) and AUPRC-Negative (PR-N) metrics, which focus specifically on the model's capacity to prioritize incorrect predictions (PR-P) or correct predictions (PR-N) in precision-recall frameworks.

Further details of implementation of prompts can be found in Appendix B, and the details of the tasks and their metrics can be found in Appendix A.

Model	GSM.	Date	Cou.	Stra.	Sport	Law	Eth.	Avg
LLaMA 3	74.8	37.8	50.2	26.6	30.6	30.9	15.0	38.0
+ours	45.2	11.8	34.6	29.3	16.4	**8.5**	26.5	24.6
+CoT	**5.0**	13.7	8.7	11.8	7.7	22.8	7.8	11.1
+ours+CoT	7.8	**1.0**	**6.5**	**8.6**	**5.4**	20.9	**5.0**	**7.9**
GPT-3.5	62.6	60.2	45.6	29.6	25.3	43.2	26.0	41.8
+ours	22.8	33.0	27.5	14.9	12.2	24.0	**10.8**	20.7
+CoT	20.3	30.8	41.8	26.0	20.5	44.3	24.8	29.8
+ours+CoT	**6.5**	**6.7**	**19.7**	**14.5**	**10.8**	**16.0**	15.1	**12.7**
GPT-4	53.5	25.7	23.7	16.8	18.7	19.4	**5.1**	23.3
+ours	27.5	19.5	18.0	13.5	13.6	11.2	9.9	16.2
+CoT	6.5	**6.6**	4.9	18.5	9.2	23.0	6.1	10.7
+ours+CoT	**4.3**	8.4	**1.6**	**11.5**	**5.6**	**9.9**	9.3	**7.2**

(a) ECE ↓ (Lower is Better)

Model	GSM.	Date	Cou.	Stra.	Sport	Law	Eth.	Avg
LLaMA 3	53.7	50.3	50.3	58.8	51.5	51.0	42.3	51.1
+ours	67.1	61.0	67.2	59.0	53.8	58.9	62.9	61.4
+CoT	55.1	54.3	50.0	64.6	74.1	54.3	54.2	58.1
+ours+CoT	**81.2**	**70.7**	**87.6**	**74.5**	**79.4**	**64.7**	**81.2**	**77.0**
GPT-3.5	55.8	56.6	50.3	53.3	52.8	51.7	54.8	53.6
+ours	82.9	60.5	**82.5**	58.6	61.5	60.6	71.0	68.2
+CoT	56.2	49.8	50.1	56.4	62.7	53.0	65.2	56.2
+ours+CoT	**85.6**	**76.0**	**82.5**	**67.5**	**66.4**	**61.5**	**86.7**	**75.2**
GPT-4	52.0	50.5	50.4	55.6	57.9	56.6	84.1	58.2
+ours	83.9	64.2	72.7	63.7	63.3	59.7	77.6	69.3
+CoT	52.1	75.1	50.0	68.8	65.0	59.5	87.6	65.5
+ours+CoT	**86.0**	**81.5**	**93.3**	**70.3**	**75.2**	**68.4**	**93.0**	**81.1**

(b) AUROC ↑ (Higher is Better)

Model	GSM.	Date	Cou.	Stra.	Sport	Law	Eth.	Avg
LLaMA 3	81.7	38.2	50.2	33.0	34.8	40.8	30.3	44.1
+ours	**88.5**	**51.7**	**66.1**	32.1	37.7	47.0	**52.3**	**53.6**
+CoT	12.4	13.4	8.7	29.7	38.3	38.2	26.1	23.8
+ours+CoT	41.3	24.0	49.0	**43.2**	**47.5**	**50.6**	51.7	43.9
GPT-3.5	76.9	**66.0**	46.1	37.6	32.3	54.8	37.5	50.2
+ours	**93.5**	64.1	**76.4**	43.5	41.7	**64.4**	51.7	**62.2**
+CoT	26.5	28.0	42.0	36.8	46.9	55.5	45.7	40.2
+ours+CoT	70.9	49.7	74.7	**48.9**	**52.2**	60.9	**75.0**	61.8
GPT-4	55.3	26.5	24.3	30.7	23.9	41.3	58.4	37.2
+ours	**83.5**	35.9	47.3	37.5	33.2	46.2	45.0	46.9
+CoT	10.5	47.6	4.9	**48.7**	27.7	46.2	**83.2**	38.4
+ours+CoT	59.3	**61.8**	**44.9**	38.9	**48.7**	**55.4**	74.9	**54.9**

(c) PR-N ↑ (Higher is Better)

Model	GSM.	Date	Cou.	Stra.	Sport	Law	Eth.	Avg
LLaMA 3	21.3	62.5	49.9	77.2	66.9	61.0	71.5	58.6
+ours	28.1	68.0	62.3	79.6	67.1	66.2	79.1	64.3
+CoT	95.4	89.3	91.3	85.2	89.2	66.5	83.0	85.7
+ours+CoT	**97.8**	**92.6**	**98.1**	**90.0**	**92.8**	**73.6**	**93.1**	**91.1**
GPT-3.5	30.0	41.5	54.5	66.2	70.8	47.5	69.6	54.3
+ours	60.4	54.4	**83.3**	71.1	**76.1**	52.3	80.2	68.3
+CoT	80.5	71.5	58.2	71.5	70.2	48.4	75.8	68.0
+ours+CoT	**92.1**	**88.4**	83.2	**77.7**	75.0	**57.3**	**91.8**	**80.8**
GPT-4	47.3	74.5	76.5	77.6	82.6	67.5	94.5	74.4
+ours	76.8	81.2	84.4	81.4	84.4	71.3	92.0	81.6
+CoT	93.7	**94.2**	95.1	81.4	88.0	68.3	92.8	87.6
+ours+CoT	**97.3**	94.0	**99.2**	**87.6**	**91.1**	**75.6**	**97.9**	**91.8**

(d) PR-P ↑ (Higher is Better)

Table 1: Comparison with Vanilla Verbalized Confidence Elicitation, while metrics (ECE, AUROC ,PR-N and PR-P) are in percentage(%). Abbreviations are used: GSM. (GSM8K), Date (Date Understanding), Cou. (Object Counting), Stra. (StrategyQA) Sport (Sport Understanding), Law (Professional Law), Eth. (Business Ethic). ECE > 0.25, AUROC, AUPRC-Positive, AUPRC-Negative < 0.6 denote significant deviation from ideal performance. The best among the same model are in bold.

5.3 Comparison with Vanilla Verbalized Confidence Elicitation

Table 1 presents comprehensive evaluations across four critical metrics: Expected Calibration Error (ECE), Area Under ROC Curve (AUROC), Precision-Recall at N (PR-N), and Precision-Recall at Precision Threshold (PR-P). Our analysis reveals three key findings. First, in the failure detection task measured by AUROC, our method demonstrates consistent superiority over vanilla models across all datasets under Chain-of-Thought (CoT) settings. Specifically, significant improvements are observed in LLaMA-3 (25.6% on Sport), GPT-3.5 (32.4% on Count), and GPT-4 (43.3% on Count) configurations. Notably, even without CoT integration, our approach maintains competitive performance - achieving 20.6% and 31.9% improvements over baselines in LLaMA-3 (Eth. dataset) and GPT-4 (GSM8K) settings respectively, though we observe a singular exception where GPT-4 baseline outperforms our method by 6.5% on Eth. dataset.

Second, detailed analysis of PR-N and PR-P metrics reveals two noteworthy patterns. While CoT integration generally enhances positive sample identification at the cost of negative class performance degradation (e.g., PR-N scores dropping from 81.7% to 12.4% in LLaMA-3/GSM8K configuration), our method effectively bridges this performance gap. Through dual optimization, we elevate PR-N scores from (81.7%, 12.4%) to (88.5%, 41.3%) and PR-P scores from (21.3%, 95.4%) to (28.1%, 97.8%) in respective settings, demonstrating balanced improvements across both metrics.

Finally, in confidence calibration measured by ECE, our method achieves state-of-the-art performance across most configurations. The most striking result emerges in LLaMA-3/Date setting where our CoT-enhanced approach attains an exception-

Method	GSM8K		Law		Date		Strategy		Ethics		Average	
	ECE	AUROC	ECE	AUROC	ECE	AUROC	ECE	AUROC	ECE	AUROC	ECE	AUROC
CoT (M=1)	10.1	54.8	39.7	52.2	23.4	57.4	22.0	59.8	30.0	56.0	25.0	56.4
CoT+Top-K (M=1)	19.6	58.5	16.7	58.9	26.1	74.2	**14.0**	61.3	**12.4**	73.3	17.8	65.2
CoT+Misleading (M=5)	8.03	88.6	18.3	59.3	20.5	67.3	21.8	61.5	17.8	71.3	17.3	69.6
CoT+Self-Random (M=5)	**6.28**	**92.7**	26.0	**65.6**	17.0	66.8	23.3	60.8	20.7	79.0	18.7	73.0
CoT+Prompt (M=5)	35.2	74.4	31.5	60.8	23.9	69.8	16.1	61.3	15.0	79.5	24.3	69.2
CoT+**ours**	6.5	85.6	**16.0**	61.5	**6.7**	**76.0**	14.5	**67.5**	15.1	**86.7**	**11.7**	**75.4**

Table 2: Comparison with Sampling Based Baselines (Consistency+CoT) on GPT-3.5 while metrics (ECE, AUROC) are in percentage(%). The best results are in bold.

Dataset	Was Dis	JS Div	Mean	Δauroc	Δ-ece
Sport	-13.83	-41.74	-14.66	-0.56	9.27
Ethics	-21.20	-72.34	-20.49	11.58	17.29
Law	-18.49	-59.99	-17.74	-1.54	15.59
Count	-11.42	-34.23	-16.93	28.57	15.10
Strategy	-15.14	-46.66	-16.94	1.57	13.20
Date	-19.27	-60.82	-25.33	0.12	26.65
GSM8K	-24.77	-52.11	-25.98	3.65	22.69

(a) *very cautious- vanilla*

Dataset	Was Dis	JS Div	Mean	Δauroc	Δ-ece
Sport	-0.27	-4.18	-0.83	-1.04	1.86
Ethics	-1.40	-10.63	-0.04	10.92	0.74
Law	-1.11	-8.13	-1.59	2.31	-0.38
Count	-0.01	-1.86	-0.13	0.37	-1.31
Strategy	-0.28	-3.47	-0.99	-1.84	1.29
Date	-0.08	-3.19	-1.22	1.97	5.23
GSM8K	-0.50	-3.34	-1.02	1.18	-0.67

(b) *cautious- vanilla*

Dataset	Was Dis	JS Div	Mean	Δauroc	Δ-ece
Sport	-0.53	-4.93	-0.47	2.53	2.99
Ethics	-1.20	-10.08	-1.79	-2.12	-2.41
Law	-0.79	-6.45	-1.33	0.98	-0.11
Count	-0.05	-2.79	-0.09	0.14	-0.42
Strategy	-0.46	-5.52	-1.02	-2.42	1.29
Date	-0.19	-4.84	-1.44	4.87	3.55
GSM8K	-0.85	-4.41	-1.58	-1.54	0.89

(c) *confident- vanilla*

Dataset	Was Dis	JS Div	Mean	Δauroc	Δ-ece
Sport	1.11	17.43	1.71	0.76	-2.33
Ethics	1.30	14.31	2.05	-0.65	-4.15
Law	1.73	19.22	3.26	1.55	-3.81
Count	0.08	2.63	0.08	-0.11	-2.28
Strategy	0.97	13.16	1.41	-0.20	-1.36
Date	-0.19	-4.75	-0.69	6.04	4.08
GSM8K	4.08	23.72	5.36	-3.05	-6.45

(d) *very confident- vanilla*

Table 3: Effects of Prompt Steering

ally low ECE of 1%. Aggregate improvements show consistent advantages: average ECE reductions of 13.4% (without CoT) and 3.2% (with CoT) for LLaMA-3, 21.1%/17.1% for GPT-3.5, and 7.1%/3.5% for GPT-4 configurations. These results collectively validate our method's effectiveness in both failure identification and confidence calibration tasks.

5.4 Comparison with Sampling Based Baselines

Table 2 systematically compares our approach with three sampling-based baselines (Misleading, Self-Random, Prompt) on GPT-3.5 and one competitive baseline (Top-K) across five reasoning-intensive benchmarks, evaluating both calibration quality (ECE) and failure detection capability (AUROC). Three pivotal observations emerge from the experimental results. First, our method achieves the lowest average ECE (11.7%) while attaining the highest average AUROC (75.4%), demonstrating superior performance on both confidence calibration and error identification task.

To be specific, on DateUnd dataset, our approach delivers breakthrough performance: attaining an ECE of 6.7% (relatively 60.8% lower than second-best) coupled with record AUROC of 76.0%. Similar dominance is observed in Biz-Ethics where we achieve the highest AUROC (86.7%) while maintaining competitive ECE (15.1%).

6 Effects of Prompt Steering

In this section, we empirically study one fundamental questions to our work: **how will the steered**

prompt steers confidence elicitation? We investigates how steered prompts influence confidence calibration through quantitative comparisons across four steering strategies (*very cautious*, *cautious*, *confident*, *very confident*) against the vanilla baseline. Table 3 presents performance variations in Wasserstein Distance (Was Dis), Jensen-Shannon Divergence (JS Div), mean confidence (Mean), AUROC (Δauroc), and (negative for better comparison) expected calibration error (Δ-ece) across seven datasets. The sign of two distances (JS Div and Was Dis) are determined by the sign their difference of mean confidence between vanilla verbalized confidence to show directions.

Conservative Steering Effects The *very cautious* strategy induces substantial reductions in confidence divergence metrics (Was Dis: -11.42 to -24.77; JS Div: -34.23 to -72.34) compared to vanilla, suggesting diminished confidence extremity. Furthermore, this conservative alignment upgrades calibration, evidenced by Δ-ece increases of +9.27% to +26.65% across datasets. However, for failure prediction tasks measured by AUROC, the conservative steered prompts don't always promote the performance, such as in Law with -1.54%. The *cautious* variant produces milder divergence reductions (Was Dis: -1.40 to -0.01; JS Div: -10.63 to -1.86) while enjoys smaller calibration quality gain (Δ-ece: -1.31 to +5.23), indicating the degree of steering actually matters.

Confidence Boosting Effects The *confident* steering strategy reveals unexpected patterns that challenge intuitive assumptions. While designed to amplify model confidence, it paradoxically reduces confidence divergence metrics (Was Dis: -0.85 to -0.05; JS Div: -10.08 to -2.79) compared to vanilla baselines in most datasets (Table 3c). While *very confident* steering strategy behaves as expected. This phenomenon indicates not only the steering direction is needed, its magnitude should be also considered.

To conclude, **prompt steering actually changes the distribution of confidences in every task**. One may fail to steer confidence if not using a proper magnitude or prompting template. And this study explains why previously the steering study part in Appendix B.2 of (Xiong et al., 2023) fails: they only use mild steering prompt.

7 Conclusion

In this work, we propose SteeringConf, a novel framework for calibrating verbalized confidence in large language models (LLMs) through systematic prompt steering and aggregation. Our empirical analysis demonstrates that explicit linguistic manipulation (e.g., "Be *very cautious* " or "Be *very confident* ") can directionally steer LLM confidence scores, challenging prior assumptions about the rigidity of confidence calibration in LLMs. By aggregating predictions and confidences across steered prompts, SteeringConf achieves state-of-the-art performance on both confidence calibration and failure detection tasks. Experiments across seven benchmarks and three LLMs (GPT-3.5, LLaMA3-70B, GPT-4) validate our method's effectiveness.

8 Limitations

This work mainly has the following limitations:

Manually Designed Steering Prompt Our currently proposed method relies on manually designed steering prompts (e.g., "*very cautious* "). The reliance on manually designed steering prompts introduces critical constraints in scalability and generalizability. While our current framework operates with a moderate steering magnitude ($K = 2$, e.g., *very cautious*, *cautious*, *vanilla*, *confident*, *very confident*), extending to larger (e.g., $K = 5$ with finer-grained directives like *extremely cautious* or *moderately confident*) would require laborious, task-specific prompt engineering. Each additional steering direction demands careful linguistic tuning to ensure semantic distinctiveness and monotonic confidence shifts. Automatic steering prompt generation or the K-free continuous steering may be possible solutions to this limitation.

Computational Inefficiency Our SteerConf method necessitates extra K forward passes through energy-intensive LLMs, directly amplifying energy consumption and associated carbon emissions. To solve this, one should adaptively prioritize high-uncertainty samples for multi-prompt inference, while using single-prompt baselines for low-uncertainty cases using some novel criteria.

References

AI@Meta. 2024. Llama 3 Model Card.

Suhana Bedi, Yutong Liu, Lucy Orr-Ewing, Dev Dash, Sanmi Koyejo, Alison Callahan, Jason A Fries, Michael Wornow, Akshay Swaminathan, Lisa Soleymani Lehmann, et al. 2024. A systematic review of testing and evaluation of healthcare applications of large language models (llms). *medRxiv*, pages 2024–04.

Kendrick Boyd, Kevin H Eng, and C David Page. 2013. Area under the precision-recall curve: point estimates and confidence intervals. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III 13*, pages 451–466. Springer.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, volume 33, pages 1877–1901. Curran Associates, Inc.

Jiuhai Chen and Jonas Mueller. 2023. Quantifying Uncertainty in Answers from any Language Model and Enhancing their Trustworthiness. *Preprint*, arXiv:2308.16175.

Long Chen, Oleg Sinavski, Jan Hünermann, Alice Karnsund, Andrew James Willmott, Danny Birch, Daniel Maund, and Jamie Shotton. 2024. Driving with llms: Fusing object-level vector modality for explainable autonomous driving. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 14093–14100. IEEE.

I.-Chun Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. 2023. FacTool: Factuality Detection in Generative AI – A Tool Augmented Framework for Multi-Task and Multi-Domain Scenarios. *Preprint*, arXiv:2307.13528.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training Verifiers to Solve Math Word Problems. *Preprint*, arXiv:2110.14168.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did Aristotle Use a Laptop? A Question Answering Benchmark with Implicit Reasoning Strategies. 9:346–361.

Ahmad Ghazal, Tilmann Rabl, Minqing Hu, Francois Raab, Meikel Poess, Alain Crolotte, and Hans-Arno Jacobsen. 2013. Bigbench: Towards an industry standard benchmark for big data analytics. In *Proceedings of the 2013 ACM SIGMOD international conference on Management of data*, pages 1197–1208.

Mario Giulianelli, Joris Baan, Wilker Aziz, Raquel Fernández, and Barbara Plank. 2023. What Comes Next? Evaluating Uncertainty in Neural Text Generators Against Human Production Variability. *Preprint*, arXiv:2305.11707.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2023. CRITIC: Large Language Models Can Self-Correct with Tool-Interactive Critiquing. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Aditya K, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory Dickinson, Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 44123–44279. Curran Associates, Inc.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How Can We Know When Language Models Know? On the Calibration of Language Models for Question Answering. 9:962–977.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 2022. Language Models (Mostly) Know What They Know. *Preprint*, arXiv:2207.05221.

Ethan Kim. 2021. Sports understanding in bigbench.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2022. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching Models to Express Their Uncertainty in Words.

Sabrina J. Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. 2022. Reducing Conversational Agents' Overconfidence Through Linguistic Calibration. 10:857–872.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

OpenAI. 2021. ChatGPT. `https://www.openai.com/gpt-3/`. Accessed: April 21, 2023.

OpenAI. 2024. GPT-4 Technical Report. *Preprint*, arXiv:2303.08774.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2463–2473. Association for Computational Linguistics.

Thomas Savage, John Wang, Robert Gallo, Abdessalem Boukil, Vishwesh Patel, Seyed Amir Ahmad Safavi-Naini, Ali Soroush, and Jonathan H Chen. 2024. Large language model uncertainty measurement and calibration for medical diagnosis and treatment. *medRxiv*, pages 2024–06.

Vaishnavi Shrivastava, Percy Liang, and Ananya Kumar. 2023. Llamas know what gpts don't show: Surrogate models for confidence estimation. *CoRR*, abs/2311.08877.

Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Lee Boyd-Graber, and Lijuan Wang. 2022. Prompting GPT-3 To Be Reliable. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*.

Sree Harsha Tanneru, Chirag Agarwal, and Himabindu Lakkaraju. 2023. Quantifying Uncertainty in Natural Language Explanations of Large Language Models. *Preprint*, arXiv:2311.03533.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5433–5442. Association for Computational Linguistics.

Yao-Hung Hubert Tsai, Walter Talbott, and Jian Zhang. 2024. Efficient Non-Parametric Uncertainty Quantification for Black-Box Large Language Models and Decision Planning. *Preprint*, arXiv:2402.00251.

Jianfeng Wang, Rong Xiao, Yandong Guo, and Lei Zhang. 2019. Learning to count objects with few exemplar annotations. *arXiv preprint arXiv:1905.07898*.

Jun Wang, Guocheng He, and Yiannis Kantaros. 2024. Safe task planning for language-instructed multi-robot systems using conformal prediction. *arXiv preprint arXiv:2402.15368*.

Ziyu Wang and Chris Holmes. 2024. On Subjective Uncertainty Quantification and Calibration in Natural Language Generation. *Preprint*, arXiv:2406.05213v1.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed 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 35 (NeurIPS)*, volume 35, pages 24824–24837.

Xinyi Wu and Zijian Wang. 2021. Data understanding in bigbench.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2023. Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. In *ICLR*.

Miao Xiong, Shen Li, Wenjie Feng, Ailin Deng, Jihai Zhang, and Bryan Hooi. 2022. Birds of a feather trust together: Knowing when to trust a classifier via adaptive neighborhood aggregation. *arXiv preprint arXiv:2211.16466*.

Fanghua Ye, Mingming Yang, Jianhui Pang, Longyue Wang, Derek F. Wong, Emine Yilmaz, Shuming Shi, and Zhaopeng Tu. 2024. Benchmarking LLMs via Uncertainty Quantification. *Preprint*, arXiv:2401.12794.

Zhuoning Yuan, Yan Yan, Milan Sonka, and Tianbao Yang. 2021. Large-scale robust deep auc maximization: A new surrogate loss and empirical studies on medical image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3040–3049.

Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. 2023. Navigating the Grey Area: How Expressions of Uncertainty and Overconfidence Affect Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5506–5524. Association for Computational Linguistics.

A Task Statement And Metric

Confidence Calibration LLMs often exhibit misaligned confidence scores, usually overestimating the reliability of their predictions. The confidence calibration task aims to improve the calibration of LLMs by aligning their confidence scores with their actual performance. In our evaluation, we use the Expected Calibration Error (ECE) (Naeini et al., 2015; Xiong et al., 2022; Yuan et al., 2021), as the calibration metric to evaluate the calibration performance of LLMs.

Suppose we have a set of input samples $\{x_i\}_{i=1}^N$, their corresponding labels $\{y_i\}_{i=1}^N$, the LLM's predictions $\{f(x_i)\}_{i=1}^N$, and the confidence scores $\{c(x_i)\}_{i=1}^N$. We divide the samples into B bins in terms of their confidence scores: $x_i \in B_b$ if $c(x_i) \in \left[\frac{b-1}{B}, \frac{b}{B}\right)$. The Expected Calibration Error (ECE) is defined as

$$\text{ECE} = \sum_{b=1}^{B} \frac{|B_b|}{N} \left|\text{acc}(B_b) - \text{conf}(B_b)\right|, \quad (9)$$

where B_b is the set of samples in the b-th bin, $|B_b|$ is the number of samples in the b-th bin, $\text{acc}(B_b) = \frac{1}{|B_b|}\sum_{x_i \in B_b} \mathbb{I}(f(x_i) = y_i)$ is the accuracy of the samples in the b-th bin, and $\text{conf}(B_b) = \frac{1}{|B_b|}\sum_{x_i \in B_b} c(x_i)$ is the average confidence score of the samples in the b-th bin. Here, $\mathbb{I}(\cdot)$ is the indicator function, which returns 1 if the event inside the parentheses occurs, and 0 otherwise.

Failure Prediction Another important task is to directly predict whether the LLM's prediction is correct or not using the confidence score. This task is often referred to as the failure prediction task. The metric used to evaluate the performance of the failure prediction task is the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) (Boyd et al., 2013). Suppose we have a set of input samples $\{x_i\}_{i=1}^N$, we can define the failure prediction task as a binary classification problem, where the input is their confidence score $c(x_i)$ and the label is the correctness of the LLM's prediction $\mathbb{I}(f(x_i) = y_i)$. The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) is defined as

$$\text{AUC-ROC} = \int_0^1 \text{TPR}(t)\, d\text{FPR}(t), \quad (10)$$

where $\text{TPR}(t)$ is the True Positive Rate at the threshold t and $\text{FPR}(t)$ is the False Positive Rate at the threshold t, which can be computed as

$$\text{TPR}(t) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(c(x_i) \geq t)\mathbb{I}(f(x_i) = y_i), \quad (11)$$

$$\text{FPR}(t) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(c(x_i) \geq t)\left(1 - \mathbb{I}(f(x_i) = y_i)\right). \quad (12)$$

B Prompts Used For Steering Confidence

In this section, we show the detailed prompt of the spectrum: {*very cautious*, *cautious*, *vanilla*, *confident*, *very confident* } as follows.

B.1 CoT Setting

very cautious Read the question, analyze step by step, provide your answer and your confidence in this answer. Note: (1) The confidence indicates how likely you think your answer will be true. (2) You are making important decisions, thus you should avoid giving a wrong answer with high confidence. (3) You should be very cautious, and tend to give low confidence on almost all of the answers. \nUse the following format to answer:\"‘Explanation: [insert step-by-step analysis here]\nAnswer and Confidence (0-100): [ONLY the {ANSWER_TYPE}; not a complete sentence], [Your confidence level, please only include the numerical number in the range of 0-100]%"‘\nOnly give me the reply according to this format, don't give me any other words.

cautious Read the question, analyze step by step, provide your answer and your confidence in this answer. Note: (1) The confidence indicates how likely you think your answer will be true. (2) You are making important decisions, thus you should avoid giving a wrong answer with high confidence. (3) You should be very cautious, and tend to give low confidence on almost all of the answers. \nUse the following format to answer:\"‘Explanation: [insert step-by-step analysis here]\nAnswer and Confidence (0-100): [ONLY the {ANSWER_TYPE}; not a complete sentence], [Your confidence level, please only include the numerical number in the range of 0-100]%"‘\nOnly give me the reply according to this format, don't give me any other words.

vanilla Read the question, analyze step by step, provide your answer and your confidence in this

answer. Note: The confidence indicates how likely you think your answer is true.\nUse the following format to answer:\n"'Explanation: [insert step-by-step analysis here]\nAnswer and Confidence (0-100): [ONLY the {ANSWER_TYPE}; not a complete sentence], [Your confidence level, please only include the numerical number in the range of 0-100]%\n"'\nOnly give me the reply according to this format, don't give me any other words.

confident Read the question, analyze step by step, provide your answer and your confidence in this answer. Note: (1) The confidence indicates how likely you think your answer will be true. (2) You are making important decisions, thus you should avoid giving a right answer with low confidence. \nUse the following format to answer:\"'Explanation: [insert step-by-step analysis here]\nAnswer and Confidence (0-100): [ONLY the {ANSWER_TYPE}; not a complete sentence], [Your confidence level, please only include the numerical number in the range of 0-100]%"'\nOnly give me the reply according to this format, don't give me any other words.

very confident Read the question, analyze step by step, provide your answer and your confidence in this answer. Note: (1) The confidence indicates how likely you think your answer will be true. (2) You are making important decisions, thus you should avoid giving a right answer with low confidence. (3) You should be very confident, and tend to give high confidence on almost all of the answers. \nUse the following format to answer:\"'Explanation: [insert step-by-step analysis here]\nAnswer and Confidence (0-100): [ONLY the {ANSWER_TYPE}; not a complete sentence], [Your confidence level, please only include the numerical number in the range of 0-100]%"'\nOnly give me the reply according to this format, don't give me any other words.

B.2 No CoT Setting

very cautious Read the question, provide your answer and your confidence in this answer. Note: (1) The confidence indicates how likely you think your answer will be true. (2) You are making important decisions, thus you should avoid giving a wrong answer with high confidence. (3) You should be very cautious, and tend to give low confidence on almost all of the answers. \nUse the following format to answer:\n"'Answer and Confidence (0-100): [ONLY the ANSWER_TYPE; not a complete sentence], [Your confidence level, please only include the numerical number in the range of 0-100]%"'\nOnly the answer and confidence, don't give me the explanation.

Table 4: Comparison between Vanilla (before - after)

Dataset	Was Dis	JS Div	Mean	Δauroc	Δ-ece
Sport	-26.42	-49.64	-28.52	-0.30	3.44
Ethics	-12.20	-36.89	-12.54	11.09	8.74
Law	-31.83	-62.52	-30.89	3.90	22.06
Count	-0.10	-3.73	-0.11	0.12	1.12
Strategy	-35.56	-55.22	-37.64	2.46	6.11
Date	-11.82	-28.24	-12.99	5.27	6.19
GSM8K	-3.61	-10.87	-2.19	10.86	2.20

cautious Read the question, provide your answer and your confidence in this answer. Note: (1) The confidence indicates how likely you think your answer will be true. (2) You are making important decisions, thus you should avoid giving a wrong answer with high confidence. \nUse the following format to answer:\n"'Answer and Confidence (0-100): [ONLY the ANSWER_TYPE; not a complete sentence], [Your confidence level, please only include the numerical number in the range of 0-100]%"'\nOnly the answer and confidence, don't give me the explanation.

vanilla Read the question, provide your answer and your confidence in this answer. Note: The confidence indicates how likely you think your answer is true.\nUse the following format to answer:\n"'Answer and Confidence (0-100): [ONLY the ANSWER_TYPE; not a complete sentence], [Your confidence level, please only include the numerical number in the range of 0-100]%"'\nOnly the answer and confidence, don't give me the explanation.

confident Read the question, provide your answer and your confidence in this answer. Note: (1) The confidence indicates how likely you think your answer will be true. (2) You are making important decisions, thus you should avoid giving a right answer with low confidence.\nUse the following format to answer:\n"'Answer and Confidence (0-100): [ONLY the ANSWER_TYPE; not a complete sentence], [Your confidence level, please only include the numerical number in the range of 0-100]%"'\nOnly the answer and confidence, don't give me the explanation.

Method	GSM8K		Prf-Law		DateUnd		StrategyQA		Biz-Ethics		Average	
	ECE	AUROC	ECE	AUROC	ECE	AUROC	ECE	AUROC	ECE	AUROC	ECE	AUROC
ours-before	25.8	84.2	11.2	57.5	23.4	58.7	23.8	56.1	15.8	66.5	20.0	64.6
ours-after	22.8	82.9	24.0	60.6	33.0	60.5	14.9	58.6	10.8	71.0	21.1	66.7

Table 5: Comparison of ours-before and ours-after

very confident Read the question, provide your answer and your confidence in this answer. Note: (1) The confidence indicates how likely you think your answer will be true. (2) You are making important decisions, thus you should avoid giving a right answer with low confidence. (3) You should be very confident, and tend to give high confidence on almost all of the answers. \nUse the following format to answer:\n"'Answer and Confidence (0-100): [ONLY the ANSWER_TYPE; not a complete sentence], [Your confidence level, please only include the numerical number in the range of 0-100]%"'\nOnly the answer and confidence, don't give me the explanation.

C Effects of Placing Confidence Elicitation Before Or After The Answer

This section studies a relatively minor question: should the confidence assigned to question or the answer in verbalized confidence elicitation? We notice one omitted fact is that every problem itself could have a difficulty level for the model. It's like that when facing a problem, **before giving an answer**, one can evaluate how difficulty this problem is, and decide whether he can answer this question. We call this confidence before answer, while all our previous setting is called confidence after answer, i.e., giving an answer and its accompanied confidence, or several answer-confidence pairs as in TopK (Tian et al., 2023).

To investigate this, we conduct two experiments, one is for vanilla GPT-3.5 with verbalized confidence before answer and verbalized confidence after answer. This is to show under vanilla setting, will this prompt modification cause confidence and performance shifting. The results are in Table 4. The other one is combined with our method with GPT-3.5, to show if our method is stable to such modification whose results are in Table 5.

Table 4 reveals nuanced performance shifts when placing confidence elicitation before answers. One can find that confidence before answer tend to produce more conservative predictions since it has much smaller mean confidences. This indicates confidence before answers could produce better confidences in vanilla verbalized confidence elicitation. However, as demonstrated in Table 5, these two placings don't make much difference in overall results for our methods: the average AUROC difference narrows to 2.1% points (64.6% vs 66.7%), while ECE values remain within 1.1% points (20.0% vs 21.1%). Domain-specific patterns emerge – for instance, confidence-after yields substantially better ECE in Business Ethics (15.8%→10.8%) but worse calibration in Professional Law (11.2%→24.0%) – no configuration demonstrates universal superiority across metrics. This indicates our method is stable to the placing of confidence elicitation.

Despite comparable aggregate performance, we ultimately adopt confidence-after elicitation for the reason that post-answer confidence naturally accommodates multi-answer scenarios such as TopK sampling. However, confidence-before elicitation presents intriguing research opportunities for difficulty estimation and refuse-to-ask mechanisms.