Quantifying the Consistency, Fidelity, and Reliability of LLM Verbalized Confidence

Anonymous ACL submission

Abstract

Large language models (LLMs) can be prompted to express their confidence in answers to a given query, referred to as *verbalized confidence*, to help users assess trustworthiness. However, its poor calibration with factual accuracy raises questions about whether it can be trusted. To address this, we formulate a set of test metrics to evaluate verbalized confidence across a large span of LLMs along three dimensions: ***Consistency***–how stable the confidence is across diverse prompts eliciting confidence in different formats, e.g., numerical scales; ***Fidelity***–is the model faithful to its own answers, e.g., more confident about them than about counterfactual answers; ***Reliability***–how well the stated confidence aligns with the answer correctness. Our findings reveal that GPT-4o, which provides the most consistent and reliable confidence, performs sub-optimally on fidelity compared to smaller models. Furthermore, all LLMs are generally most confident about their original answers, even compared to higher-quality gold responses. Reliability is highly sensitive to the prompt format and the chosen calibration metric. Thus, we conclude that each evaluation dimension captures a distinct aspect of model trustworthiness. [1]

1 Introduction

Large Language Models (LLMs) are increasingly effective at answering fact-based questions across domains such as medicine, law, and education (Alqahtani et al., 2023; Alfertshofer et al., 2024; Xiao et al., 2025). In these high-stakes settings, *verbalized confidence* (Kadavath et al., 2023; Tian et al., 2023; Chen and Mueller, 2024), i.e., explicit expressions of how certain a model is about its answer, can improve user trust and support better human-AI interactions (Kadavath et al., 2022; Zhang et al., 2024).

Yet, despite its growing use, the trustworthiness of verbalized confidence remains an open question. Prior studies have found that these expressions are often poorly aligned with actual correctness (Groot and Valdenegro Toro, 2024; Ni et al., 2025), and sensitive to prompt design and dataset variations (Xiong et al., 2024; Xia et al., 2025). However, these works primarily highlight specific limitations without offering a systematic understanding of when and why LLMs' confidence can be trusted.

For a more comprehensive investigation, this work designs a novel evaluation setting to study the *consistency*, *fidelity*, and *reliability* of verbalized confidence in LLMs (Fig. 1). ***Consistency*** measures if an LLM expresses similar greedy-encoded confidence for the same answer when prompted differently. For example, the model is considered inconsistent if it outputs different confidence with "Provide the probability" and "Provide the confidence" prompts. We use the standard deviation to quantify the confidence consistency. ***Fidelity*** investigates how faithful LLMs are to their own answers compared to alternatives such as target, counterfactual and abstain (e.g., "I don't know") responses. For example, one would expect a faithful LLM to decrease confidence when its original answer "Berlin" is replaced with a counterfactual answer "1950" for the query "what's the capital of Germany?". We quantify fidelity by the frequency with which LLMs assign higher confidence to their own answers than to the counterfactual ones. ***Reliability*** evaluates how well the expressed confidence aligns with actual answer correctness across diverse prompts. Four established calibration metrics (e.g., Brier Score (Brier, 1950), ECE (Guo et al., 2017)) are used to represent this quality.

Using these three test settings, we evaluate the confidences elicited from 13 LLMs in five model families (GPT, Mistral, Llama, Qwen, OLMo) on five question answering datasets and ten prompting strategies. We summarize our main findings as

[1] Code is available at https://anonymous.4open.science/r/evaluation_of_confidence-34AC

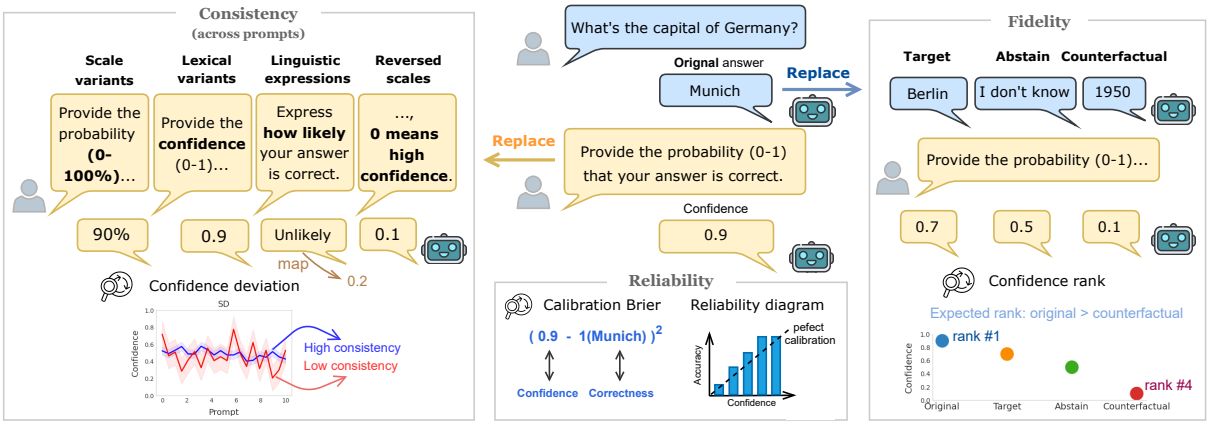


Figure 1: The design of our novel test metrics to assess LLM confidence. We evaluate *consistency* through the standard deviation (SD) of confidence scores elicited by various prompts (a lower SD indicates higher consistency). *Fidelity* checks the confidence rank of the LLM's original answer and alternative answers; a faithful LLM is expected to assign a higher confidence to its original answer than the replaced counterfactual answer. We measure *reliability* with standard calibration metrics (e.g., Brier score) and the reliability diagram.

follows:

(i) Bigger $\neq$ (always) better: Confidences elicited from larger models (e.g., GPT-4o) are more consistent and reliable than those from small models, but are not necessarily more faithful. For example, Qwen-14B outperforms GPT-4o in fidelity.

(ii) Prompt format matters: Prompts with only the change of numerical scales or lexical words (stemming from "confidence") elicit more consistent and faithful confidence than prompts that contain linguistic expressions (e.g., "Unlikely") or reversed-scale ("0 indicates high confidence"). Meanwhile, reliability depends strongly on both prompt format and the chosen calibration metric, e.g., reversed-scale prompts improve reliability only for small models.

(iii) Bias toward own outputs: All models give the highest average confidence to their original answers—even compared to replacing them with the target ground-truth. However, their confidences in other alternative responses vary widely: some models (e.g., OLMo) even elicit high confidence (>0.5) for counterfactual answers.

(iv) Model family differences: OLMo models show lower consistency and fidelity overall. In contrast, Qwen models are more consistent and faithful but less reliable. While post-training improves consistency for OLMo, it does not reliably improve the other two metrics.

These results show that even state-of-the-art language models struggle to provide verbalized confidences that are fully consistent, faithful, and reliable. Our proposed novel test suite provides a rigorous, multi-dimensional evaluation of LLM confidence estimation—assessing not just calibration, but also how consistently and faithfully models express certainty across diverse prompt types; our evaluation attributes are also general and can be applied to other types of confidence estimation methods in the future. Based on our comprehensive analysis, we recommend not using prompts with linguistic expressions and reversed scales as a practical way to generate more consistent and faithful verbalized confidence. Furthermore, this comprehensive perspective of trustworthiness is essential for developing LLMs that users can genuinely trust.

2 Related Work

Verbalized confidence in LLMs (Kadavath et al., 2022; Tian et al., 2023; Chen et al., 2024) has emerged as a promising interface for improving user trust and interpretability. However, recent work (Groot and Valdenegro Toro, 2024; Ni et al., 2025; Zhou et al., 2024) shows that verbalized confidence is often overconfident and poorly calibrated with factual correctness, indicating a need for more thorough analysis of its trustworthiness. However, Tian et al. (2023); Xiong et al. (2024); Geng et al. (2024a) solely focus on evaluating the calibration of confidence estimation. In contrast, we propose a comprehensive multi-dimensional test to assess consistency, fidelity, and reliability.

Consistency in LLM generation is often used to quantify the uncertainty of LLMs (Wang et al., 2023) by prompting the model multiple times for

the same input with a loose temperature (Kuhn et al., 2023; Xiong et al., 2024; Chen and Mueller, 2024). The certainty is calculated based on the frequency of different outputs. However, our test assesses the consistency of greedy-encoded confidence scores elicited by diverse prompts. Consistency measures the stability of all scores rather than the frequency of individual scores. **Fidelity** defined in Zhang et al. (2024) is under a multiple-choice setting and assessing an LLM's fidelity by checking if the LLM continues to select the same choice even when the content of the choice is replaced with "All other options are wrong". Their fidelity is used to improve confidence calibration. Differently, our paper evaluates the fidelity of confidence in an open-ended QA setting and examines specific confidence ranks between LLMs' original answers and substituted answers, e.g., counterfactual answers. **Reliability** of confidence is often measured by its calibration with correctness (Guo et al., 2017; Zhu et al., 2023; Huang et al., 2024; Geng et al., 2024b). Following previous work, we integrate four classic calibration metrics in our test. However, our test considers prompt sensitivity of LLMs (Zhuo et al., 2024; Errica et al., 2025; Xia et al., 2025) and therefore designed novel prompt strategies such as scaling confidence scores or reversing the scale. Such out-of-distribution prompts have never been explored in reliability evaluations. Our test brings additional insights on the sensitivity of calibration evaluation to new prompt structures.

3 Methodology

We investigate whether verbalized confidences of LLMs are trustworthy and useful for interpreting model responses based on three qualities: *consistency*, *faithfulness*, and *reliability*. To answer this, we first surface model confidence via a two-stage prompting strategy. The first-stage prompt requests the LLM to provide an answer to the input question, while the second-stage prompt elicits verbalized confidence regarding the generated answer from the LLM. Fig. 1 illustrates our approach, including the prompt design and the evaluation process.

3.1 Consistency of Confidence across Prompts

To ensure a robust consistency evaluation, we design ten variations of the second-stage prompt to elicit verbalized confidences from the LLM (full prompts are shown in Table 4 in Appendix). These prompts are categorized into four distinct types:

Scale variants. This category investigates whether LLMs can express consistent confidence scores using different numerical scales. We employ three prompts to ask probability in (1) 0 to 1 scale (**P(1)**), (2) 0% to 100% scale (**P(%)**), and (3) 0 to 10 scale (**P(10)**). A consistent LLM should provide confidence of 90% and 9 for P(%) and P(10) if its confidence for P(1) is 0.9.

Lexical variants. We use three prompts that are semantically equivalent but vary in the lexical key term used to elicit a numerical score between 0 and 1: (1)"probability" (**P(1)**), (2) "certainty" (**CT(1)**), and (3) "confidence" (**CF(1)**). This allows us to assess whether models treat these interchangeable terms consistently when quantifying their confidence with the same score across prompts.

Linguistic expressions. Following Tian et al. (2023), we include the same prompt for LLMs to express certainty within a list of 13 pre-defined linguistic expressions (**L.**) from "Almost No Chance" to "Almost certain". We also extend the prompt by converting the expression list into a multiple-choice format, assigning a unique alphabetical label to each expression (**L. MC**) such as "a: Almost No Chance". This evaluates whether the model consistently maps linguistic expressions to their corresponding choices. A consistent LLM should select the choice "a" for the L. MC prompt if it answers "Almost No Chance" for the L. prompt. Those selected expressions are converted to numerical scores according to rules in Appendix A.1.

Reversed scales. To evaluate whether the LLM can flexibly adapt to alternate representations of confidence, we reverse the orientation of the numerical confidence scales such that a score of 0 now indicates the highest confidence, and higher values represent lower confidence. We replicate the three scales from the scale category, resulting in three reverse prompts: **RP(1)**, **RP(%)**, and **RP(10)**. We expect an LLM to perform relatively complex in-context reasoning here to generate a consistent confidence, even though these prompts are highly likely out of distribution for the training data.

Confidence scores generated by using different prompts are then normalized to the same scale (0-1) for comparison and analysis.

3.1.1 Consistency Evaluation

We use the **mean standard deviation** (MSD, ↓) to quantify consistency. MSD first gets the standard deviation of confidence scores provided by prompts for each data sample and then takes the mean value

of all samples. The formula for calculating the consistency in a group of k prompts over n data samples is:

$$MSD = \frac{1}{n}\sum_{i=1}^{n}\sqrt{\frac{1}{k-1}\sum_{j=1}^{k}(c_{ij}-\bar{c_i})^2} \quad (1)$$

Where c_{ij} denotes the confidence for the data sample i generated using the j-th prompt. Additionally, we measure the confidence correlation strength between prompt pairs using **pearson correlation coefficient**.

3.2 Fidelity of Confidence

We assess the fidelity of confidence scores by analyzing how confidence shifts under controlled perturbations to their original generated answer. This setting tests an LLM's confidence sensitivity to changes in the response and evaluates whether the LLM can generate reasonable confidence estimates across answers of varying correctness. We consider four experimental settings:

Original: The LLM is queried for its confidence in its original (unmodified) answer. This is the default setting for verbalized confidence evaluations. **Target**: We replace the LLM's original answer with the target (ground-truth) answer and measure the confidence assigned to this revised response. **Abstain**: We replace the original answer with an explicit abstention ("I don't know the answer") and record the corresponding confidence score. **Counterfactual**: The original answer is replaced with a randomly sampled, incorrect response from the same dataset. We then measure the LLM's confidence in this counterfactual answer.

3.2.1 Fidelity Evaluation

The confidence rank of the original and replaced answers is accessed by their mean confidence across all data samples. We expect that a faithful LLM should rank its original answer's confidence higher than that of a counterfactual answer. Therefore, we defined **Fidelity rate (F, ↑)** that measures the percentage of samples where the confidence of the original answers is higher than that of counterfactual answers, formally:

$$F = \frac{1}{n}\sum_{i=0}^{n}\mathbf{1}(c_{ij}^{o} > c_{ij}^{c}) \quad (2)$$

Where c_{ij}^{o} and c_{ij}^{c} denote the confidence of **o**riginal and **c**ounterfactual answers using the j-th prompt on data sample i.

3.3 Reliability of Confidence

Reliability of the LLM confidence measures how well-calibrated it is with the correctness of the associated answer (Guo et al., 2017). A reliable LLM should offer confidence aligned with the correctness of its outputs. For assessing the correctness of an open-ended answer given by an LLM, we follow Xia et al. (2025) to employ a Judge model, Prometheus-8x7b-v2.0 (Kim et al., 2024), to decide the semantic equivalence between the generated answer and the target answer.

3.3.1 Reliability Evaluation

We utilize four classic calibration metrics used in Ulmer et al. (2024) to evaluate confidence reliability, which are the Brier score (Brier, 1950), ECE (Guo et al., 2017), SMECE (Błasiok and Nakkiran, 2024) and AUROC. We report the Brier results in the main papers and others in Appendix A.6.

Brier (↓). An LLM achieves a low calibration Brier score if c_i accurately reflects the reliability of the generated answer a_i, which is the binary correctness $\mathbf{1}(a_i)$. Specifically, for all n generated answers, the Brier score is the average squared error between all predicted confidence and the correctness of these answers:

$$Brier = \frac{1}{n}\sum_{i=1}^{n}(c_i - \mathbf{1}(a_i))^2 \quad (3)$$

Reliability diagram. This diagram is a visualization that compares an LLM's confidence with its actual accuracy. All confidence scores are grouped into confidence bins, and for each bin, the average confidence is plotted against the observed accuracy. A perfectly calibrated LLM will have points that lie on the diagonal, when accuracy matches the confidence.

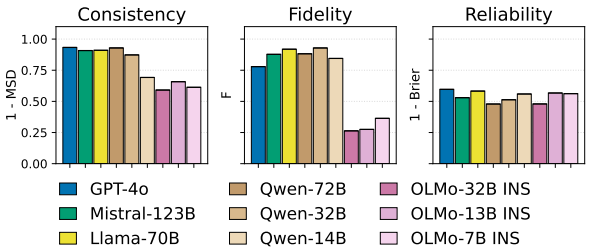
4 Experimental Setup

Our experiments are conducted on an aggregation of five question answering datasets and 13 LLMs across different model families, sizes, and post-training schemes. We apply greedy decoding to all LLMs across experiments to ensure consistent generated outputs.

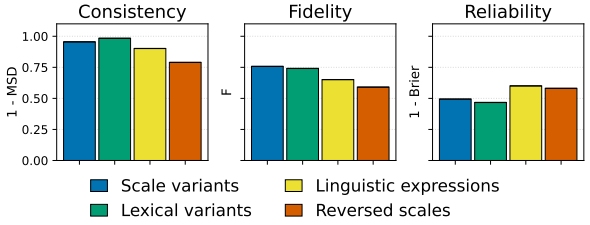
Datasets. We perform our confidence verbalization experiments on a combined dataset of five open-ended question-answering benchmarks covering diverse topics. It includes 1,000 randomly sampled examples each from Natural Questions (NQ) (Kwiatkowski et al., 2019), SciQ (Johannes Welbl,

2017), TrivialQA (Joshi et al., 2017), and PopQA (Mallen et al., 2023), along with the full TruthfulQA (Lin et al., 2022) set (817 examples), totaling 4,817 questions.

LLMs. We consider 13 LLMs from five model families: (1). **GPT** (OpenAI, 2024): GPT-4o (GPT-4o-2014-11-20) ; (2). **Mistral** (Mistral AI, 2024): Mistral-123B (Mistral-Large-Instruct-2411); (3). **Llama** (Meta AI, 2024): Llama-70B (Llama-3.3-70B-Instruct); (4). **Qwen** models across different model sizes (Yang et al., 2024): Qwen-72B&-32B&-14B (Qwen2.5-72B&-32B&-14B-Instruct); (5). **OLMo** models across different sizes and post-training schemes (Team OLMo et al., 2024): OLMo2-32B&13B&7B (OLMo-2-0325-32B-Instruct, OLMo-2-1124-13B&7B-Instruct), OLMo-13B-SFT&DPO (OLMo-2-1124-13B-sft&-dpo), OLMo-7B-SFT&DPO (OLMo-2-1124-7B-sft&-dpo).



(a) Mean fidelity and reliability of all instruction models over all prompts. Consistency is measured on all prompts (same values as the **All** column in Table 1).



(b) Prompt formatting matters: Mean influences of prompt types on fidelity and reliability evaluation and further averaged over all instruction models. Consistency is measured on prompts in each type and averaged over all instruction models.

Figure 2: Overview of the main evaluation results.

5 Results and Analysis

We summarize the overall results in Fig. 2. The upper figure shows that GPT-4o, which performs the best in consistency and reliability, falls short in fidelity. The lower figure demonstrates that prompt formatting can impact the evaluation, e.g., using scale-reversed prompts can decrease LLMs' consistency and fidelity. We analyze more detailed results in the following sections.

Model	Sca.	Lex.	Lin.	Rev.	All
GPT-4o	0.038	0.017	0.049	0.056	0.067
Mistral-123B	0.032	0.011	0.091	0.039	0.092
Llama-70B	0.065	0.025	0.041	0.047	0.090
Qwen-72B	0.066	0.014	0.031	0.057	0.071
Qwen-32B	0.040	0.008	0.049	0.151	0.127
Qwen-14B	0.051	0.015	0.092	0.403	0.308
OLMo-32B INS	0.007	0.006	0.081	0.564	0.409
OLMo-13B INS	0.065	0.020	0.051	0.456	0.342
OLMo-13B DPO	0.064	0.022	0.039	0.345	0.344
OLMo-13B SFT	0.064	0.026	0.098	0.548	0.356
OLMo-7B INS	0.040	0.030	0.400	0.120	0.387
OLMo-7B DPO	0.063	0.037	0.360	0.152	0.380
OLMo-7B SFT	0.097	0.076	0.413	0.190	0.381
Avg.	0.053	0.024	0.138	0.241	0.258

Table 1: *Consistency* results measured with MSD (↓, mean standard deviation) across **all** prompts and prompts in scale variants (**Sca.**), lexical variants (**Lex.**) , linguistic expressions (**Lin.**) and reversed scales (**Rev.**) categories. A darker color indicates better consistency performance.

5.1 How Consistent are LLMs' Confidence across Prompts?

Table 1 presents an overview of confidence consistency evaluated with different prompting formatting. We also examine the pairwise correlations of confidences between prompts for individual models (Fig. 3 is for OLMo models, and other models are shown by Fig. 7-10 in the Appendix). We found that larger models maintain higher consistency of confidence, showing their capability of responding to complex confidence-eliciting prompts.

LLMs remain consistent with the change of scales and lexical words in prompts. LLMs get lower MSD scores on prompts with only scale and lexical changes compared to adding linguistic expressions or reversed scales. Small models (<= 32B) struggle more in maintaining consistency, especially with the reversed scale prompts. For example, Fig. 3 demonstrates the low correlation of RP(1) and RP(%) prompts to other prompts on different sizes of OLMo models.

Larger models are generally more consistent. The consistency of all models in Table 1 demonstrates that larger models (>70B parameters) provide more consistent confidence than smaller models. Similarly, scores of Qwen models show that consistency is correlated with model size–larger models are more consistent. However, this does not apply to OLMo models, where OLMo-32B INS is less consistent than both OLMo-13B and -7B models. We hypothesize that OLMo models can vary

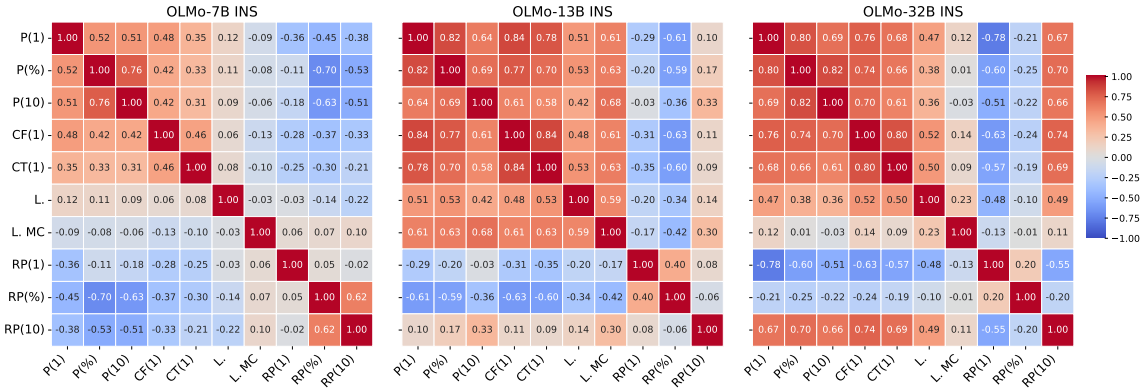


Figure 3: *Consistency*: Pearson correlations of prompts in confidence on OLMo models with different sizes.

Model	P(1)	P(%)	P(10)	CF(1)	CT(1)	L.	L. MC	RP(1)	RP(%)	RP(10)	Avg.
GPT-4o	0.759	0.786	0.779	0.837	0.767	0.829	0.796	0.831	0.620	0.786	0.779
Mistral-123B	0.871	0.877	0.881	0.873	0.855	0.918	0.920	0.870	0.817	0.894	0.878
Llama-70B	0.900	0.863	0.897	0.945	0.923	0.946	0.933	0.956	0.934	0.890	0.919
Qwen-72B	0.874	0.900	0.933	0.889	0.874	0.922	0.664	0.895	0.930	0.940	0.882
Qwen-32B	0.928	0.919	0.936	0.956	0.943	0.931	0.948	0.813	0.954	0.960	0.929
Qwen-14B	0.936	0.944	0.938	0.891	0.907	0.934	0.942	0.705	0.337	0.913	0.845
OLMo-32B INS	0.389	0.345	0.256	0.360	0.298	0.315	0.062	0.019	0.001	0.590	0.263
OLMo-13B INS	0.401	0.419	0.563	0.347	0.359	0.257	0.230	0.040	0.091	0.044	0.275
OLMo-13B DPO	0.375	0.458	0.583	0.355	0.334	0.211	0.117	0.034	0.133	0.096	0.270
OLMo-13B SFT	0.444	0.041	0.372	0.281	0.410	0.184	0.303	0.040	0.039	0.002	0.212
OLMo-7B INS	0.742	0.724	0.694	0.572	0.613	0.160	0.016	0.035	0.071	0.011	0.364
OLMo-7B DPO	0.707	0.755	0.646	0.624	0.624	0.245	0.041	0.061	0.070	0.024	0.380
OLMo-7B SFT	0.787	0.779	0.718	0.741	0.735	0.245	0.019	0.036	0.048	0.028	0.414
Avg.	0.701	0.678	0.707	0.667	0.665	0.546	0.461	0.410	0.388	0.475	

Table 2: *Fidelity* results measured with **F** (↑, fidelity rate). A darker color indicates better performance.

in performance due to their different post-training data, which may impact their complex reasoning capability.

Post-training schemes improve consistency within each prompt category. The result of OLMo models shows that post-training schemes such as DPO (i.e., direct preference optimization) and RLVR (i.e., reinforcement learning with verified reward) improve the LLM's consistency from the SFT (supervised fine-tuning) models within each prompt category (e.g., scale variants), but this trend does not always apply to the consistency over all prompts.

5.2 How Faithful are LLMs' Confidence?

Table 2 shows the fidelity scores of all models using different prompts. We observe that the SOTA model (GPT-4o) displays lower fidelity than smaller models. Fig. 4, which illustrates the confidence ranks of different substituted answers, shows that LLMs give the highest average confidence to their original answer, followed by the target answer.

Prompts with only scale changes elicit faithful confidences the best. We observe that P(1), P(%) and P(10) achieve higher fidelity scores than other prompts. This demonstrates that fidelity is robust to scale variants in prompts. In contrast, reversed-scale prompts get the lowest average fidelity score compared to other prompts. E.g., RP(%) gets the lowest score and even decreases the score of GPT-4o by more than 10% from P(1) prompt. Similarly, the fidelity score of Qwen-14B decreases by 60% when using RP(%). This indicates that LLMs are not robust to reverse scale, possibly because such instruction is not seen in the pre-/post-training.

Fidelity does not correlate with model size. We find that large models such as GPT-4o get lower fidelity scores than smaller models like Llama and Qwen. Similarly, Qwen-32B achieves a better score than Qwen-72B. We also notice that OLMo models consistently show lower fidelity scores compared to other model families, regardless of size or post-training methods. Notably, Qwen-14B achieves over 50% higher fidelity than OLMo-32B

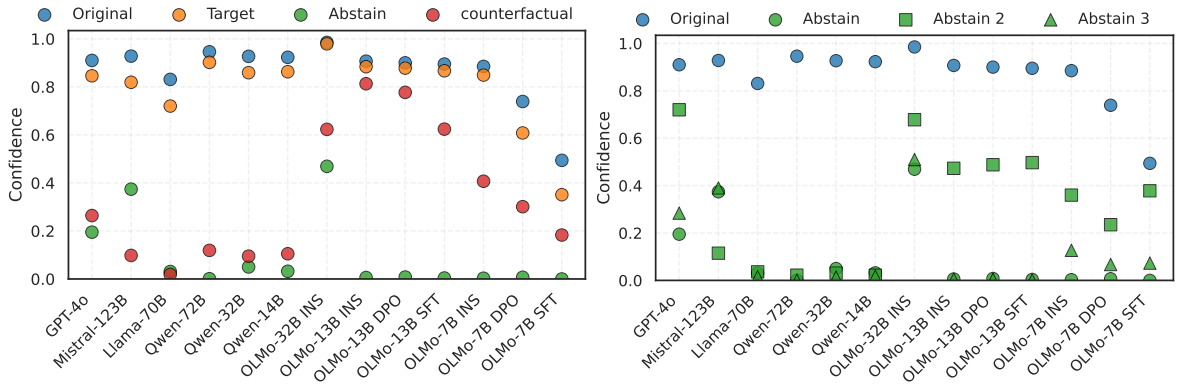


Figure 4: *Fidelity*: (left) The mean confidence of original, target, abstain and counterfactual answers. (Right) The mean confidence of different abstain answers. Abstain 2 ("I cannot be sure about my answer") and Abstain 3 ("That's out of my current knowledge base") serve different interpretations of Abstain–"I don't know the answer".

Model	P(1)	P(%)	P(10)	CF(1)	CT(1)	L.	L. MC	RP(1)	RP(%)	RP(10)	Avg.
GPT-4o	0.411	0.422	0.387	0.409	0.411	0.369	0.401	0.403	0.448	0.383	0.404
Mistral-123B	0.492	0.480	0.468	0.510	0.511	0.412	0.338	0.508	0.523	0.473	0.471
Llama-70B	0.422	0.412	0.380	0.418	0.431	0.400	0.400	0.442	0.462	0.422	0.419
Qwen-72B	0.550	0.530	0.423	0.542	0.543	0.485	0.524	0.570	0.566	0.487	0.522
Qwen-32B	0.552	0.546	0.478	0.536	0.536	0.426	0.393	0.473	0.518	0.435	0.489
Qwen-14B	0.567	0.547	0.471	0.550	0.554	0.446	0.372	0.282	0.305	0.395	0.449
OLMo-32B INS	0.614	0.624	0.625	0.611	0.610	0.493	0.412	0.380	0.374	0.599	0.534
OLMo-13B INS	0.549	0.534	0.452	0.557	0.583	0.327	0.293	0.279	0.285	0.565	0.442
OLMo-13B DPO	0.538	0.517	0.441	0.553	0.580	0.330	0.312	0.301	0.293	0.383	0.425
OLMo-13B SFT	0.537	0.629	0.545	0.550	0.584	0.344	0.257	0.346	0.354	0.640	0.479
OLMo-7B INS	0.573	0.592	0.554	0.688	0.657	0.228	0.471	0.256	0.255	0.229	0.450
OLMo-7B DPO	0.561	0.565	0.534	0.692	0.647	0.213	0.429	0.243	0.272	0.219	0.437
OLMo-7B SFT	0.520	0.520	0.500	0.677	0.595	0.225	0.468	0.264	0.316	0.217	0.430
Avg.	0.530	0.532	0.481	0.561	0.557	0.361	0.390	0.365	0.382	0.419	

Table 3: *Reliability* results measured with Brier score(↓). A darker color indicates better performance.

INS. Furthermore, post-training techniques do not appear to improve fidelity: both OLMo-7B DPO and INS perform worse than the base OLMo-7B INS model.

Why LLMs have different confidence ranks for the abstain answer? Fig. 4 (left) shows that most LLMs tend to rank the abstain answer lower than the counterfactual answer. However, the Mistral and Llama models show an opposite trend. Specifically, Mistral gives a mean confidence of 0.4 to abstain answers, which contrasts with the confidence below 0.1 given by most other LLMs. We hypothesize that the abstain answer—"I don't know the answer" might be interpreted differently by LLMs, based on their different training schemes. For example, it can indicate that LLMs receive uncertainty about this answer and therefore give a neutral confidence, or it can be interpreted as a lack of knowledge about the question and thus imply a low confidence.

To further explain this, we apply this setting with two more explicit abstain answers: "I cannot be sure about the answer" (Abstain 2) and "That's outside my current knowledge base" (Abstain 3). Fig. 4 (right) shows close average confidence of Abstain and Abstain 3 across models, implying that **"I don't know the answer" is more likely interpreted as a lack of knowledge for LLMs.** This potentially explains why most LLMs display low confidence for the Abstain answer. Meanwhile, GPT-4o and OLMo models increase by more than 20% of confidence for the Abstain 2 setting, suggesting that model fidelity across answers can be low even when alternative answers are semantically equivalent.

5.3 How Reliable are LLMs' Confidence?

Table 3 presents the calibration Brier scores of LLMs for all prompts. We observe that all LLMs demonstrate high Brier scores, which validates that

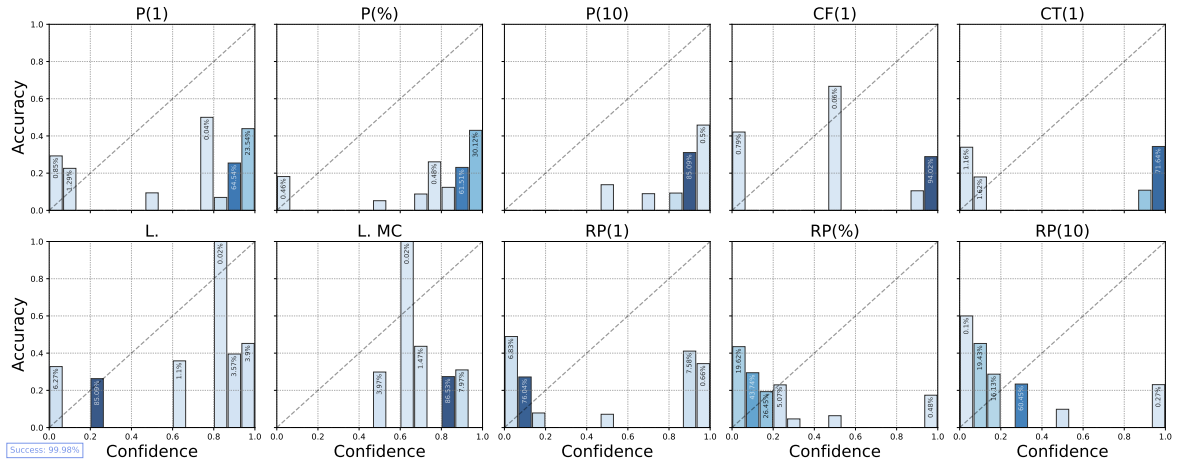


Figure 5: *Reliability:* Reliability diagrams of OLMo-7B INS across all confidence-elicit prompts. A darker bin color indicates a higher percentage of the data samples falling into the bin.

LLMs struggle to express well-calibrated verbalized confidence (Ni et al., 2025). Results of all four calibration metrics (the other three metrics are shown by Table 5-7 in Appendix) show that the **calibration of verbalized confidence highly depends on the prompts and the evaluated metric.** Specifically, large models like GPT-4o and Llama-70B are relatively more stable with a difference of around 0.08 across prompts, while smaller models like Qwen-14B or OLMo-7B, the difference can reach 0.46. This shows the high risk of getting unreliable confidence when using an inappropriate prompt. Besides, the observations are also impacted by the evaluation metric. For example, prompts with linguistic expressions and reversed scales perform better on three calibration metrics, i.e., ECE and SMECE, but fall short in AUROC (Table 5, 6, and 7 in Appendix).

Discussion of calibration metrics. We observe that smaller LLMs, such as OLMo-7B INS, often achieve lower Brier, ECE, and SMECE scores when prompted with scale-reversed prompts (RP). However, these lower scores are largely attributed to the models' incapability of following such prompts, leading them to generate disproportionately low confidence scores (shown by Fig. 5). Given that small LLMs typically have lower task accuracy than large models, this tendency to produce low confidence inadvertently improves their calibration. Crucially, this does not imply that smaller models are better calibrated or that RP prompts systematically improve calibration. In fact, when smaller models exhibit higher accuracy on a task, their calibration scores can degrade under the same prompts. We also find that Brier, ECE, and SMECE tend to produce highly correlated results across models and prompts, yet they often diverge from AUROC. This discrepancy arises because Brier etc. emphasize the absolute confidence value, and reward if a low-accuracy model generally gives low confidence. In contrast, AUROC measures whether more reliable answers receive higher confidence, independent of whether the absolute confidence values correspond to the accuracy.

6 Conclusion

In this paper, we present a comprehensive set of metrics for evaluating the confidence of LLMs along three key dimensions: *consistency*, *fidelity*, and *reliability*. Through a comprehensive analysis of 13 LLMs across five QA datasets and ten diverse prompting strategies, we find that no state-of-the-art model excels across all dimensions. This highlights a fundamental limitation of the existing one-dimensional evaluation approach, i.e., calibration metrics, which fail to capture important aspects of model confidence. For instance, while GPT-4o achieves the highest calibration scores, it underperforms in fidelity compared to smaller models, illustrating a disconnect between calibration and meaningful self-assessment. We also demonstrate that prompt formatting impacts confidence evaluation and recommend prompts without reversed scale or linguistic expressions to elicit more consistent and faithful confidence. Finally, our proposed test metrics are general, making them applicable to more confidence estimation methods, offering a foundation for future research into improving and interpreting confidence estimation in broader model behaviors.

Limitations

Our paper mainly focuses on evaluating verbalized confidence—that is, confidence explicitly expressed by language models in natural language form. While this type of confidence is especially important for interpretability and human-facing applications, it represents only one class of confidence estimation methods. Other approaches, such as logit-based, embedding-based, or Bayesian uncertainty measures, are also valuable for assessing model confidence and could offer complementary insights. Evaluating these alternative methods under similar multi-dimensional criteria (e.g., consistency, fidelity, reliability) would be a promising direction for future work, especially in understanding how different confidence signals align or diverge across model architectures and task settings.

References

Michael Alfertshofer, Cosima C Hoch, Paul F Funk, Katharina Hollmann, Barbara Wollenberg, Samuel Knoedler, and Leonard Knoedler. 2024. Sailing the seven seas: a multinational comparison of chatgpt's performance on medical licensing examinations. *Annals of Biomedical Engineering*, 52(6):1542–1545.

Tariq Alqahtani, Hisham A Badreldin, Mohammed Alrashed, Abdulrahman I Alshaya, Sahar S Alghamdi, Khalid Bin Saleh, Shuroug A Alowais, Omar A Alshaya, Ishrat Rahman, Majed S Al Yami, et al. 2023. The emergent role of artificial intelligence, natural learning processing, and large language models in higher education and research. *Research in social and administrative pharmacy*, 19(8):1236–1242.

Jarosław Błasiok and Preetum Nakkiran. 2024. Smooth ECE: Principled reliability diagrams via kernel smoothing. In *The Twelfth International Conference on Learning Representations*.

Glenn W. Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1 – 3.

Jiuhai Chen and Jonas Mueller. 2024. Quantifying uncertainty in answers from any language model and enhancing their trustworthiness. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5186–5200, Bangkok, Thailand. Association for Computational Linguistics.

Meiqi Chen, Yixin Cao, Yan Zhang, and Chaochao Lu. 2024. Quantifying and mitigating unimodal biases in multimodal large language models: A causal perspective. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16449–16469, Miami, Florida, USA. Association for Computational Linguistics.

Federico Errica, Davide Sanvito, Giuseppe Siracusano, and Roberto Bifulco. 2025. What did I do wrong? quantifying LLMs' sensitivity and consistency to prompt engineering. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1543–1558, Albuquerque, New Mexico. Association for Computational Linguistics.

Wade Fagen-Ulmschneider. 2023. Perception of probability words. Manuscript, University of Illinois Urbana-Champaign.

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

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2024b. A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6577–6595.

Tobias Groot and Matias Valdenegro Toro. 2024. Overconfidence is key: Verbalized uncertainty evaluation in large language and vision-language models. In *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024)*, pages 145–171, Mexico City, Mexico. Association for Computational Linguistics.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1321–1330. JMLR.org.

Yukun Huang, Yixin Liu, Raghuveer Thirukovalluru, Arman Cohan, and Bhuwan Dhingra. 2024. Calibrating long-form generations from large language models. *Preprint*, arXiv:2402.06544.

Matt Gardner Johannes Welbl, Nelson F. Liu. 2017. Crowdsourcing multiple choice science questions. *arXiv:1707.06209v1*.

Mandar Joshi, Eunsol Choi, Daniel S 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.

Chethan Kadavath, Amanda Askell, Jackson Kernion, Tom Henighan, Ben Mann, Gretchen Krueger, Sarah Kreps, Aaron McKane, Gaurav Mistry, Joe Kim, et al.

2023. Prompting GPT-3 to be reliable. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*.

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.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. *Preprint*, arXiv:2405.01535.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. 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.

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.

Meta AI. 2024. Introducing meta llama 3: The most capable openly available llm to date. Accessed: 2025-05-01.

Mistral AI. 2024. Mistral large instruct 2411. Accessed: 2025-05-01.

Shiyu Ni, Keping Bi, Lulu Yu, and Jiafeng Guo. 2025. Are large language models more honest in their probabilistic or verbalized confidence? In *Information Retrieval*, pages 124–135, Singapore. Springer Nature Singapore.

OpenAI. 2024. Hello gpt-4o. Accessed: 2025-05-01.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2024. 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*.

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*, pages 5433–5442, Singapore. Association for Computational Linguistics.

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

vLLM Contributors. 2023. vllm: A high-throughput and memory-efficient llm serving library. `https://github.com/vllm-project/vllm`. Accessed: 2025-01.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*.

Yuxi Xia, Pedro Henrique Luz de Araujo, Klim Zaporojets, and Benjamin Roth. 2025. Influences on llm calibration: A study of response agreement, loss functions, and prompt styles. *Preprint*, arXiv:2501.03991.

Dongfu Xiao, Chen Gao, Zhengquan Luo, Chi Liu, and Sheng Shen. 2025. Can llms assist computer education? an empirical case study of deepseek. *Preprint*, arXiv:2504.00421.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Mozhi Zhang, Mianqiu Huang, Rundong Shi, Linsen Guo, Chong Peng, Peng Yan, Yaqian Zhou, and Xipeng Qiu. 2024. Calibrating the confidence of large language models by eliciting fidelity. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2959–2979, Miami, Florida, USA. Association for Computational Linguistics.

Kaitlyn Zhou, Jena Hwang, Xiang Ren, and Maarten Sap. 2024. Relying on the unreliable: The impact of language models' reluctance to express uncertainty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3623–3643, Bangkok, Thailand. Association for Computational Linguistics.

Chiwei Zhu, Benfeng Xu, Quan Wang, Yongdong Zhang, and Zhendong Mao. 2023. On the calibration of large language models and alignment. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9778–9795.

Jingming Zhuo, Songyang Zhang, Xinyu Fang, Haodong Duan, Dahua Lin, and Kai Chen. 2024. ProSA: Assessing and understanding the prompt sensitivity of LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1950–1976, Miami, Florida, USA. Association for Computational Linguistics.

A Appendix

A.1 Prompt Details

Below is a standard prompt template used in our experiments for prompting LLMs. The second stage prompt to elicit confidence can be placed by prompts in Table 4.

> *User:* Answer the question, give ONLY the answer, no other words or explanation: [question]
> *Assistant:* [answer]
> *User:* Provide the probability that your answer is correct. Give ONLY the probability between 0.0 and 1.0, no other words or explanation.

For the linguistic prompts, we followed Tian et al. (2023) to map the linguistic expressions to numerical scores. The specific mapping are ["Almost No Chance", "Highly Unlikely", "Chances are Slight", "Little Chance", "Unlikely", "Probably Not", "About Even", "Better than Even", "Likely", "Probably", "Very Good Chance", "Highly Likely", 'Almost Certain"] to [0.02, 0.05, 0.1, 0.1, 0.2, 0.25, 0.5, 0.6, 0.7, 0.7, 0.8, 0.9, 0.95] according to Fagen-Ulmschneider (2023).

A.2 Technical Details

We use vLLM (vLLM Contributors, 2023) library for LLM inference and serving, the temperature of all LLMs is set to 0 for greedy-encoding model outputs. All our experiments are conducted on NVIDIA HGX H100, which requires approximately 200 GPU hours to replicate.

A.3 Accuracy of Models

We present the detailed accuracy of models on each benchmark dataset in Fig. 6.

A.4 Consistency: prompt confidence correlation

Similar to Figure 3 in the main paper, the prompt correlations of other models are shown by Figure 7-10.

A.5 Fidelity: confidence differences and distribution

Similar to Fig. 4, we investigate more detailed confidence differences for original right and wrong answers provided by LLMs. The results are shown in Fig. 11 and 12. We observe a very similar trend to the overall confidence difference shown in Fig. 4.

Furthermore, we demonstrate the detailed confidence distribution of all substituted answers of different models in Fig. 13.

A.6 Reliability: results of more evaluation metrics and reliability diagram

To provide a robust evaluation of reliability, we followed (Ulmer et al., 2024) and tested all the prompts and models with three more calibration metrics: ECE, SMECE and AUROC. The results are shown by Table 5, 6 and 7.

We also report the reliability diagrams of all instruction models in Fig. 14 - 19.

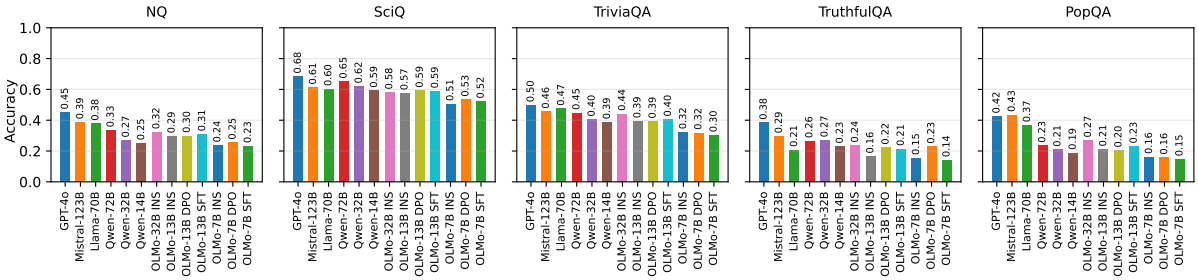


Figure 6: Accuracy of models on different datasets.

Concept	Prompt Name	Prompt
Numerical scale	P(1)	Provide the probability that your answer is correct. Give ONLY the probability **between 0.0 and 1.0**, no other words or explanation.
	P(%)	Provide the probability that your answer is correct. Give ONLY the probability **between 0% and 100%**, no other words or explanation.
	P(10)	Provide the probability that your answer is correct. Give ONLY the probability **between 0 and 10**, no other words or explanation.
Lexical variations	CF(1)	Provide the **confidence** that your answer is correct. Give ONLY the confidence between 0.0 and 1.0, no other words or explanation.
	CT(1)	Provide the **certainty** that your answer is correct. Give ONLY the confidence between 0 and 10, no other words or explanation.
Linguistic expressions	L.	Describe how likely it is that your answer is correct as **one of the following expressions**: ['Almost No Chance', 'Highly Unlikely', 'Chances are Slight', 'Little Chance', 'Unlikely', 'Probably Not', 'About Even', 'Better than Even', 'Likely', 'Probably', 'Very Good Chance','Highly Likely', 'Almost Certain']. Give ONLY the chosen expression, no other words or explanation.
	L. MC	Describe how likely it is that your answer is correct by choosing **one of the following options**: [a: 'Almost No Chance', b: 'Highly Unlikely', c: 'Chances are Slight', d: 'Little Chance', e:'Unlikely', f: 'Probably Not', g: 'About Even', h: 'Better than Even', i: 'Likely', j: 'Probably', k; 'Very Good Chance', l: 'Highly Likely', m: 'Almost Certain']. Give ONLY the chosen option, no other words or explanation.
Reverse scale	AP(1)	Provide the probability that your answer is correct. Give ONLY the probability between 0.0 and 1.0 (**where 0.0 means that the answer is completely correct, 1.0 means completely wrong**), no other words or explanation.
	AP(%)	Provide the probability that your answer is correct. Give ONLY the probability between 0% and 100% (**where 0% means that the answer is completely correct, 100% means completely wrong**), no other words or explanation.
	AP(10)	Provide the probability that your answer is correct. Give ONLY the probability between 0 and 10 (**where 0 means that the answer is completely correct, 10 means completely wrong**), no other words or explanation.

Table 4: Detail of prompts for eliciting confidence from LLMs.

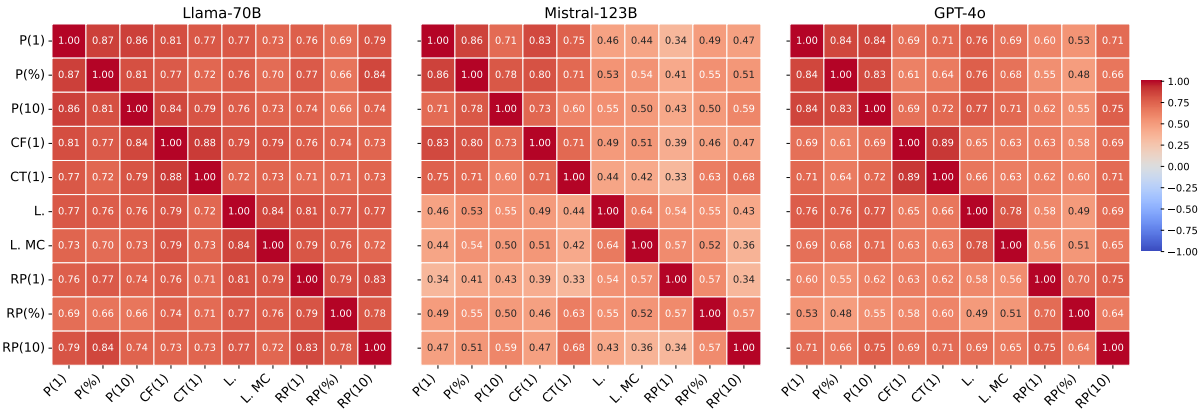


Figure 7: Consistency: Pearson correlations of prompts in confidence on large models from different model families

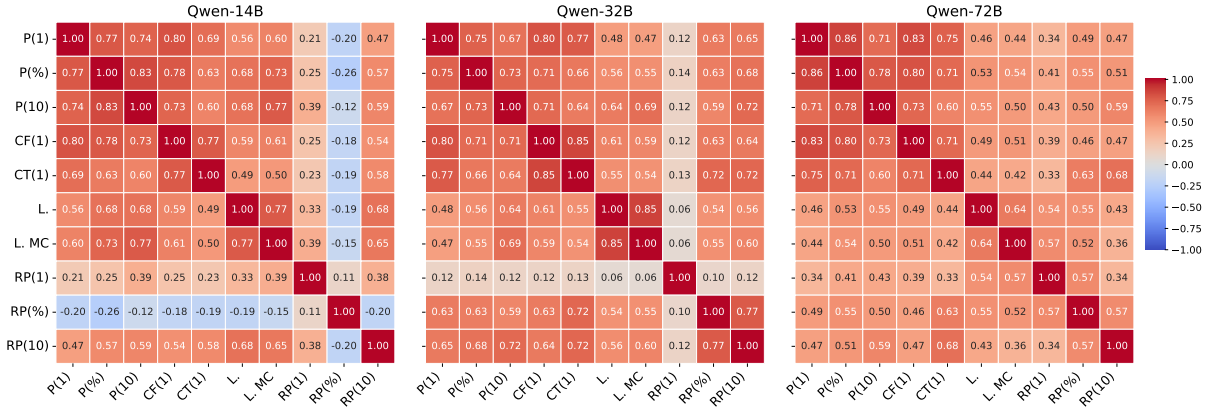


Figure 8: Consistency: Pearson correlations of prompts in confidence on Qwen models with different sizes.

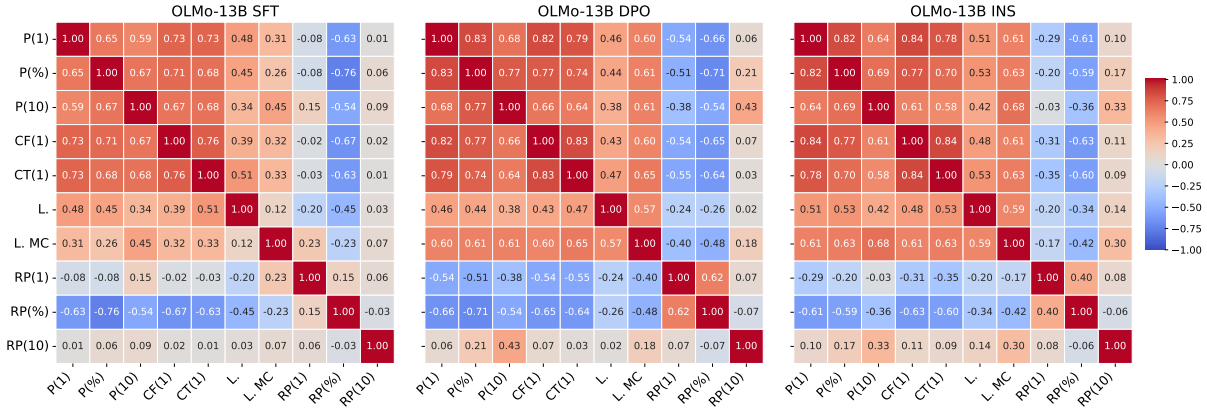


Figure 9: Consistency: Pearson correlations of prompts in confidence on OLMo-13B models with different fine-tuning schemes.

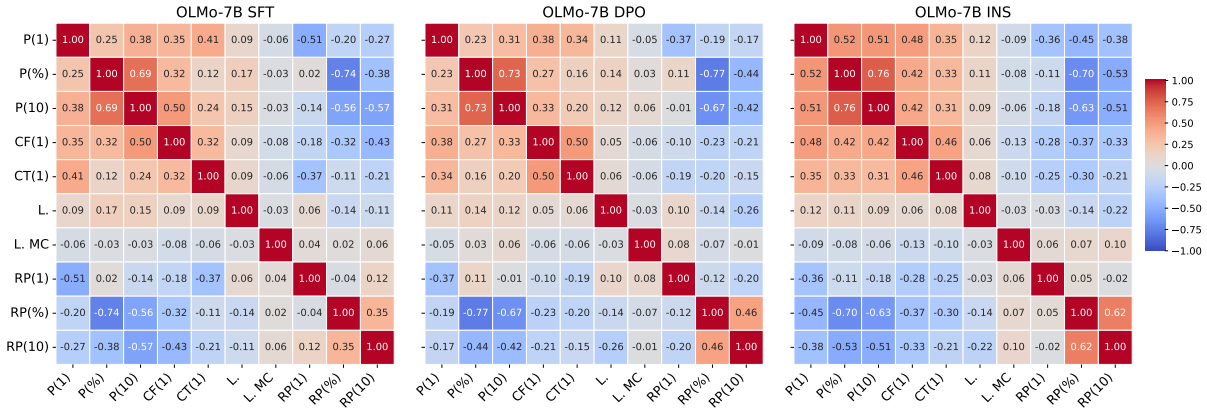


Figure 10: Consistency: Pearson correlations of prompts in confidence on OLMo-7B models with different fine-tuning schemes.

Model	P(1)	P(%)	P(10)	CF(1)	CT(1)	L.	L. MC	RP(1)	RP(%)	RP(10)	Avg.
GPT-4o	0.426	0.436	0.401	0.423	0.427	0.360	0.409	0.414	0.452	0.397	0.414
Mistral-123B	0.498	0.487	0.479	0.518	0.519	0.418	0.321	0.512	0.527	0.485	0.477
Llama-70B	0.438	0.426	0.396	0.444	0.454	0.404	0.411	0.456	0.474	0.433	0.434
Qwen-72B	0.564	0.548	0.456	0.558	0.559	0.501	0.535	0.576	0.571	0.523	0.539
Qwen-32B	0.572	0.573	0.522	0.562	0.562	0.458	0.429	0.503	0.543	0.475	0.520
Qwen-14B	0.591	0.580	0.520	0.582	0.585	0.482	0.411	0.269	0.298	0.438	0.476
OLMo-32B INS	0.617	0.624	0.625	0.616	0.614	0.508	0.422	0.381	0.374	0.608	0.539
OLMo-13B INS	0.581	0.570	0.505	0.587	0.609	0.322	0.292	0.239	0.254	0.574	0.453
OLMo-13B DPO	0.566	0.553	0.488	0.579	0.603	0.314	0.308	0.261	0.237	0.362	0.427
OLMo-13B SFT	0.562	0.635	0.576	0.574	0.604	0.335	0.218	0.313	0.352	0.640	0.481
OLMo-7B INS	0.612	0.630	0.597	0.699	0.675	0.112	0.514	0.219	0.192	0.139	0.439
OLMo-7B DPO	0.605	0.613	0.583	0.705	0.671	0.086	0.461	0.208	0.259	0.129	0.432
OLMo-7B SFT	0.556	0.569	0.553	0.692	0.620	0.120	0.509	0.230	0.334	0.064	0.425
Avg.	0.553	0.557	0.515	0.580	0.577	0.340	0.403	0.352	0.374	0.405	

Table 5: Reliability results measured with ECE(↓). A darker color indicates better performance.

Model	P(1)	P(%)	P(10)	CF(1)	CT(1)	L.	L. MC	RP(1)	RP(%)	RP(10)	Avg.
GPT-4o	0.305	0.314	0.277	0.309	0.295	0.314	0.347	0.281	0.230	0.261	0.293
Mistral-123B	0.270	0.289	0.284	0.293	0.277	0.351	0.292	0.277	0.266	0.307	0.291
Llama-70B	0.338	0.296	0.326	0.357	0.351	0.343	0.349	0.275	0.263	0.251	0.315
Qwen-72B	0.401	0.404	0.362	0.409	0.393	0.395	0.412	0.303	0.279	0.337	0.370
Qwen-32B	0.429	0.418	0.403	0.423	0.420	0.374	0.358	0.363	0.414	0.382	0.399
Qwen-14B	0.439	0.432	0.398	0.430	0.423	0.387	0.349	0.182	0.238	0.312	0.359
OLMo-32B INS	0.290	0.287	0.287	0.295	0.288	0.398	0.355	0.186	0.187	0.315	0.289
OLMo-13B INS	0.431	0.399	0.352	0.434	0.433	0.292	0.267	0.213	0.214	0.279	0.332
OLMo-13B DPO	0.422	0.392	0.347	0.430	0.426	0.285	0.274	0.243	0.214	0.202	0.324
OLMo-13B SFT	0.416	0.306	0.332	0.424	0.402	0.292	0.206	0.203	0.177	0.293	0.305
OLMo-7B INS	0.445	0.419	0.442	0.485	0.417	0.111	0.403	0.188	0.129	0.097	0.314
OLMo-7B DPO	0.435	0.402	0.435	0.484	0.456	0.086	0.376	0.167	0.127	0.083	0.305
OLMo-7B SFT	0.414	0.375	0.421	0.481	0.442	0.121	0.401	0.189	0.172	0.047	0.306
Avg.	0.387	0.364	0.359	0.404	0.387	0.288	0.338	0.236	0.224	0.244	

Table 6: Reliability results measured with SMECE(↓). A darker color indicates better performance.

Model	P(1)	P(%)	P(10)	CF(1)	CT(1)	L.	L. MC	RP(1)	RP(%)	RP(10)	Avg.
GPT-4o	0.666	0.681	0.674	0.668	0.665	0.673	0.617	0.677	0.573	0.675	0.657
Mistral-123B	0.586	0.614	0.615	0.586	0.573	0.664	0.601	0.571	0.545	0.618	0.597
Llama-70B	0.702	0.694	0.706	0.715	0.724	0.691	0.644	0.668	0.635	0.664	0.684
Qwen-72B	0.654	0.630	0.707	0.631	0.677	0.675	0.529	0.578	0.550	0.702	0.633
Qwen-32B	0.593	0.679	0.729	0.641	0.652	0.754	0.718	0.605	0.616	0.727	0.671
Qwen-14B	0.639	0.642	0.736	0.676	0.706	0.732	0.726	0.694	0.425	0.760	0.673
OLMo-32B INS	0.515	0.500	0.499	0.526	0.519	0.520	0.516	0.473	0.500	0.570	0.514
OLMo-13B INS	0.672	0.681	0.706	0.668	0.689	0.547	0.640	0.505	0.415	0.568	0.609
OLMo-13B DPO	0.634	0.689	0.692	0.641	0.670	0.531	0.590	0.456	0.349	0.646	0.590
OLMo-13B SFT	0.614	0.528	0.624	0.616	0.660	0.470	0.651	0.605	0.496	0.503	0.577
OLMo-7B INS	0.627	0.627	0.570	0.565	0.624	0.521	0.498	0.458	0.376	0.394	0.526
OLMo-7B DPO	0.633	0.672	0.603	0.563	0.642	0.540	0.513	0.469	0.340	0.411	0.538
OLMo-7B SFT	0.609	0.683	0.663	0.587	0.638	0.547	0.503	0.515	0.341	0.458	0.554
Avg.	0.626	0.640	0.656	0.622	0.649	0.605	0.596	0.559	0.474	0.592	

Table 7: Reliability results measured with AUROC (↑). A darker color indicates better performance.

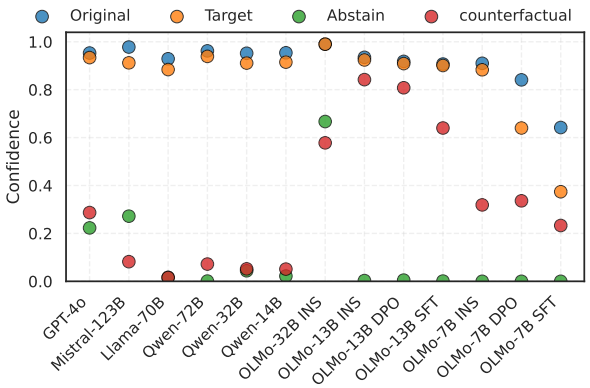


Figure 11: *Fedility*: The mean confidence of original, target, abstain and counterfactual answers for original right answers.

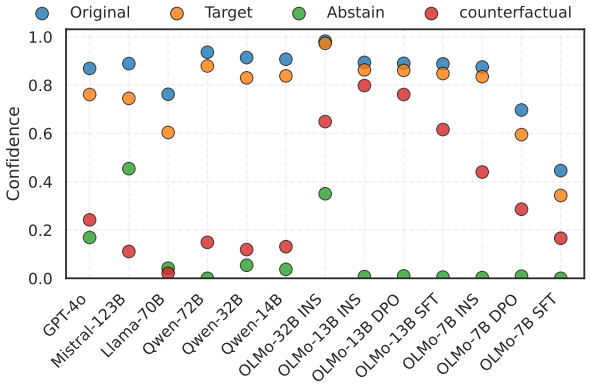


Figure 12: *Fedility*: The mean confidence of original, target, abstain and counterfactual answers for original wrong answers.

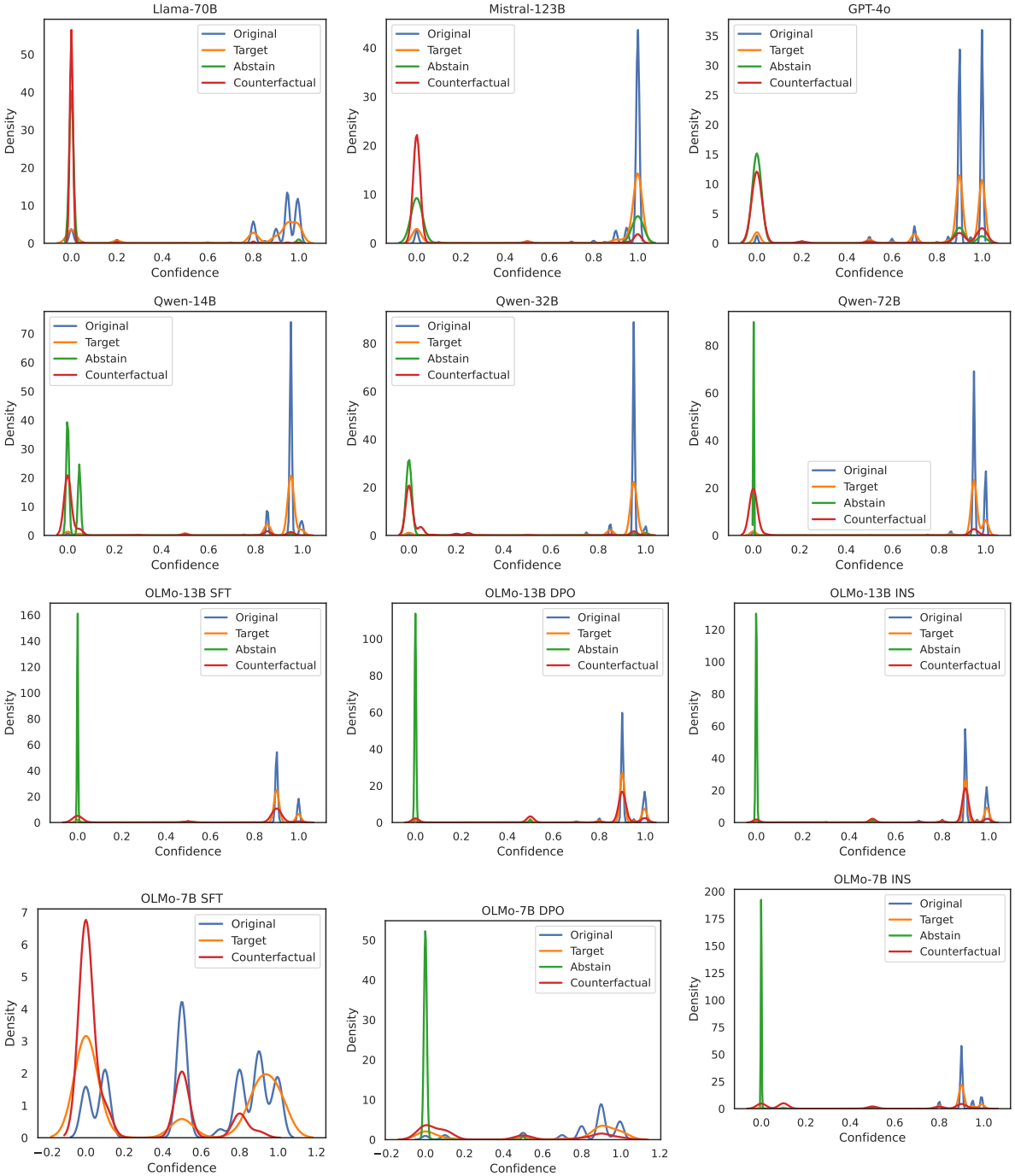


Figure 13: *Fedility*: Confidence score distribution of original, target, abstain and counterfactual answers.

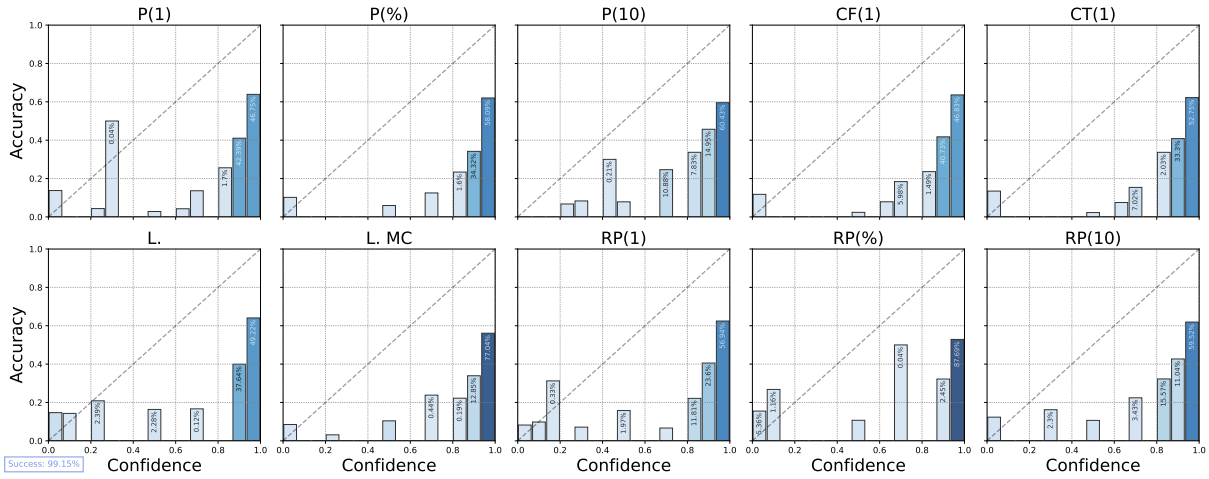


Figure 14: *Reliability:* Reliability diagrams of GPT-4o across all confidence-elicit prompts. A darker bin color indicates a higher percentage of the data samples falling into the bin.

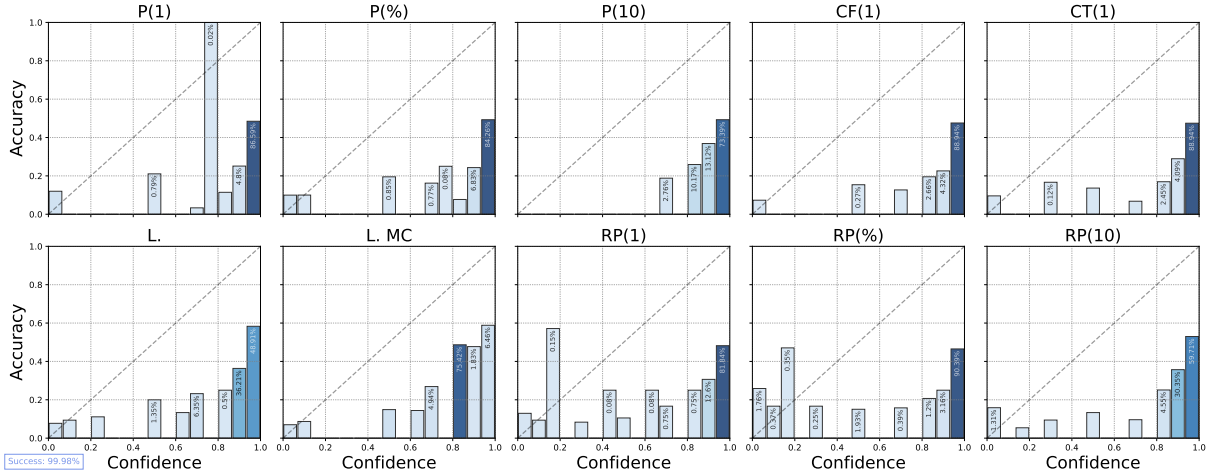


Figure 15: *Reliability*: Reliability diagrams of Mistral-123B across all confidence-elicit prompts. A darker bin color indicates a higher percentage of the data samples falling into the bin.

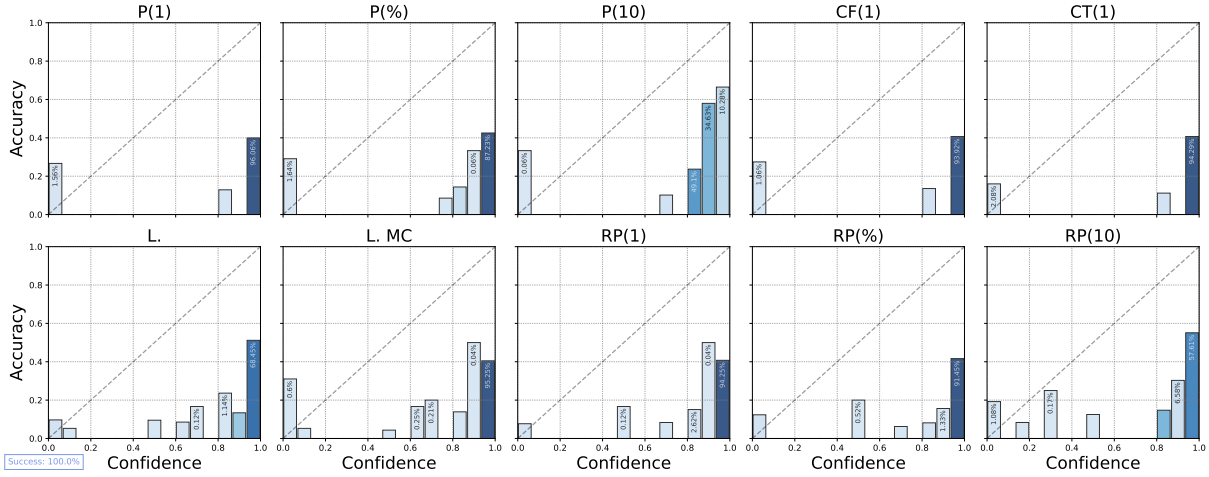


Figure 16: *Reliability*: Reliability diagrams of Qwen-72B across all confidence-elicit prompts. A darker bin color indicates a higher percentage of the data samples falling into the bin.

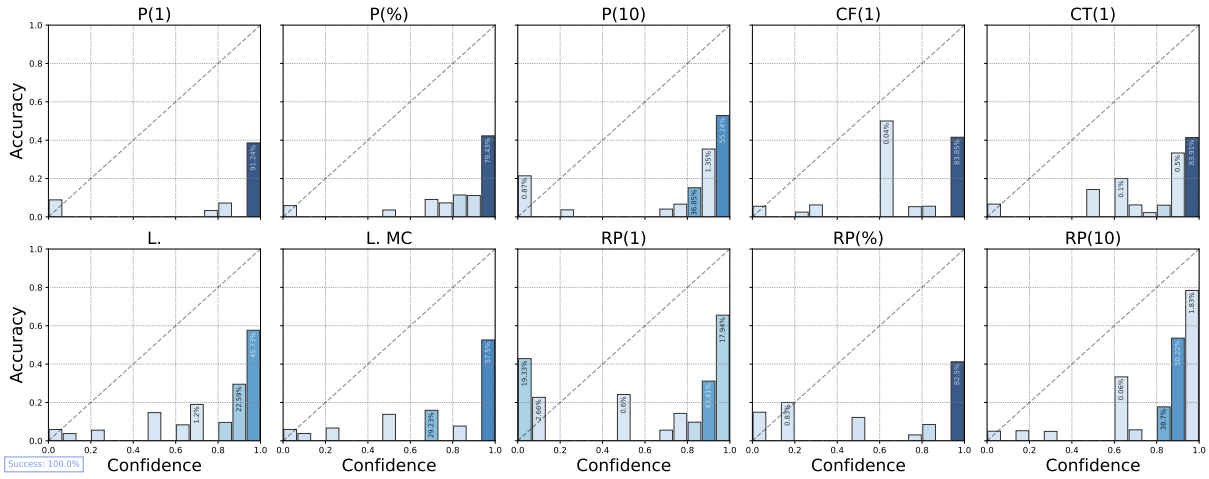


Figure 17: *Reliability*: Reliability diagrams of Qwen-32B across all confidence-elicit prompts. A darker bin color indicates a higher percentage of the data samples falling into the bin.

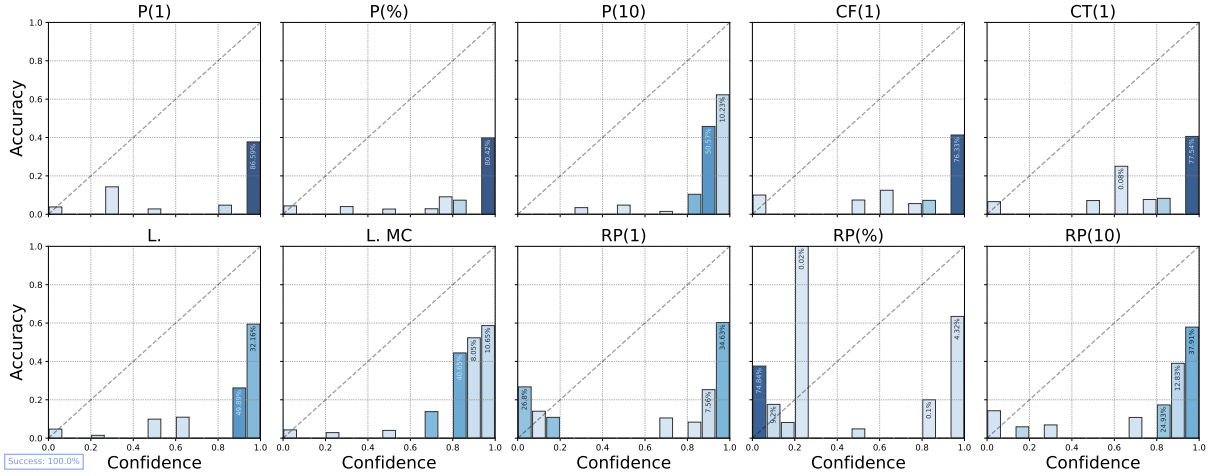


Figure 18: *Reliability*: Reliability diagrams of Qwen-14B across all confidence-elicit prompts. A darker bin color indicates a higher percentage of the data samples falling into the bin.

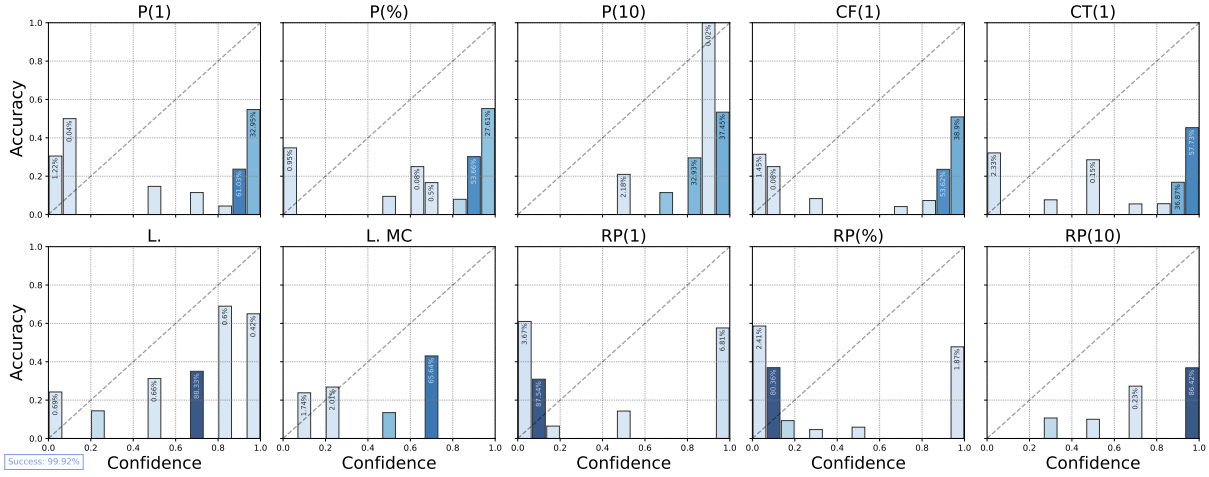


Figure 19: *Reliability*: Reliability diagrams of OLMo-13B INS across all confidence-elicit prompts. A darker bin color indicates a higher percentage of the data samples falling into the bin.