Uncertainty Estimation of Large Language Model Replies in Natural Sciences

Philip Müller[1]Ꝑ, Nicholas Popovič[2], Michael Färber[2], Peter Steinbach[1]

1 HelmholÀ-Zentrum Dresden-Rossendorf, Group of Artificial Intelligence 2 Technische Universität Dresden, Scalable Software Architectures for Data Analytics

Motivation

Large Language Models (LLMs) hold significant potential for integration into scientific research workflows, supporting tasks such as literature review, knowledge comprehension, and data analysis to improve efficiency. However, their adoption is hindered by a critical issue: LLMs can generate confident-sounding but factually incorrect responses, so-called hallucinations, which are often difficult to detect. The need for continuous manual verification of outputs would largely negate the anticipated efficiency gains. To address this challenge, automated **Uncertainty Quantification (UQ)** is essential. UQ methods can help identify low-confidence or potentially erroneous outputs, enabling a more responsible use of LLM generations in downstream scientific tasks. This research aims to develop a reproducible benchmarking framework for evaluating uncertainty metrics, focusing on the calibration of token-level probabilities and assessing the reliability of various sequence-level UQ metrics in the context of scientific question answering.

Summary/ Results

In this research we presents the first large-scale benchmark of UQ metrics for LLMs in scientific question answering. It addresses key gaps in the literature by evaluating the calibration of token probabilities and comparing multiple sequence-level uncertainty metrics.

Regarding the token level calibration, we assessed the impact of instruction tuning on calibration. Across datasets of varying reasoning complexity, instruction tuning was found to induce polarization, where models became overconfident regardless of answer correctness. This greatly reduces the reliability of token probabilities as uncertainty indicators.

Assessing sequence level calibration, we benchmarked four selected uncertainty metrics across eight datasets. We find that Verbalization-based metrics are no reliable proxies for uncertainty estimation in scientific question answering. The Frequency of Answer metric, which uses semantic consistency across multiple generations, aligned beŁer with correctness but requires high computational effort and methods of clustering answers semantically. Projecting this approach to token level, as is promised by Claim-Conditioned Probability, fails due to vanishing confidence scores with generations lengths and unreliable methods for idenficiation of semantic equivalence between tokens.

Finally, the LM-Polygraph framework [2] was reengineered into a modular, extensible platform for scalable, reproducible benchmarking. These contributions lay the groundwork for future research and development of uncertainty quantification in scientific question answering.

Implementation of Benchmarking Framework

We build upon the LM-Polygraph framework by Fadeeva et al. [2], which provides a robust foundation for computing a diverse range of uncertainty metrics. In this framework, the calculation logic is modularized into intermediate and final steps, each encapsulated in dedicated classes with clearly defined dependencies. To support scalable and reproducible benchmarking, we extensively reworked the framework. While LM-Polygraph already introduced modularity, our implementation generalizes this approach by representing each step as a configurable node. An acyclic directed execution graph is dynamically constructed at runtime based on the specified configuration. This structure enables optimizations such as asynchronous processing and efficient reuse of cached results. A layered execution wrapper manages resource allocation, batching, the skipping of redundant computations, and persistent caching using disk-backed stores. The result is an extensible and efficient framework for reproducibly benchmarking large datasets using time-intensive resources.

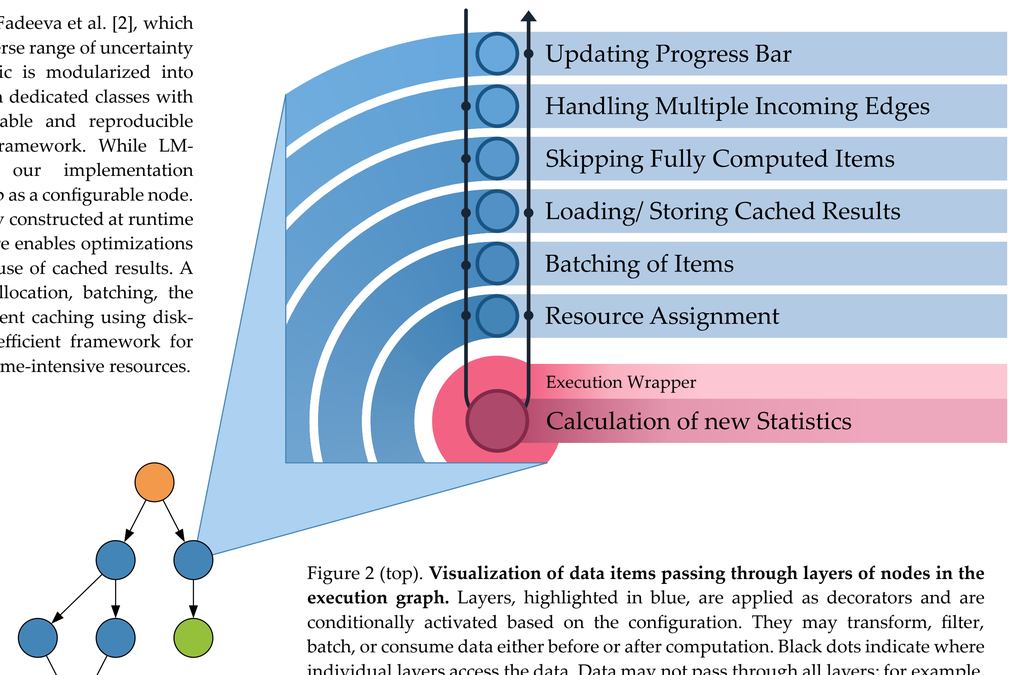


Figure 2 (top). **Visualization of data items passing through layers of nodes in the execution graph.** Layers, highlighted in blue, are applied as decorators and are conditionally activated based on the configuration. They may transform, filter, batch, or consume data either before or after computation. Black dots indicate where individual layers access the data. Data may not pass through all layers; for example, if a node's result is cached for an item, the Caching Layer retrieves and returns the cached result without invoking lower layers. The execution wrapper manages the invocation of the computation encapsulated by the node.

Figure 1 (right). **Abstract Visualization of the Order of Computation Steps as Acyclic Directed Graph in Reworked LM-Polygraph.** This features the dataset (orange), intermediate computation steps (blue) and calculation of the final uncertainty metric scores (green).

Token Level Calibration

Previous Work

In the GPT-4 Technical Report [1], OpenAI compared the calibration of responses from base model and instruction-tuned GPT-4. They used the multiple choice dataset MMLU, rephrasing prompts as classification tasks, where the model is to pick the answer choice it considers most correct by responding with the label A, B, C or D representing the possible answer choices. The probabilities assigned to the labels are used as confidence in the answer choices. The results suggested good calibration in the base model (ECE 0.007), but significantly worse calibration in the instruction-tuned model (ECE 0.074). This sparked a controversy about the influence of instruction tuning on the calibration of models. The hypothesis was that due to training the model on additional instruction data during instruction tuning, the resulting token probability distributions would no longer represent the original training data, which is considered the ground truth.

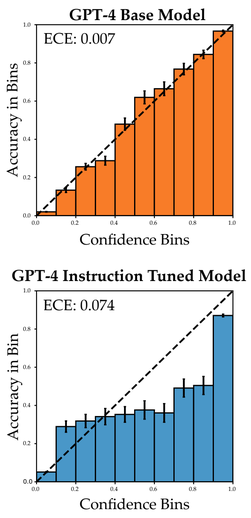


Figure 3. **Calibration plots of the base model (top) and the instruction tuned (boÇom) GPT-4 model on a subset of the MMLU dataset.** On the x-axis are bins according to the model's confidence (logprob) in each of the A/B/C/D choices for each question; on the y-axis is the accuracy within each bin. The doLed diagonal line represents perfect calibration. Right: Calibration plot of the post-trained GPT-4 model on the same subset of MMLU. The post-training hurts calibration significantly. **Adapted from the GPT-4 Technical Report [1]**

Experiment Design

To assess token-level calibration and the effect of instruction tuning, we expanded on the experiment from the GPT-4 Technical Report [1]. Four multiple-choice datasets with different reasoning demands (MMLU, ArcReasoning, GSM8K, GPQA) were evaluated using three model size pairs (7 B, 24 B, 70 B), each comprising a base model and its instruction-tuned counterpart. We compared four 3-shot prompt designs to query the label of the answer the model considers most correct as a single-token answer. We selected the best prompt based on the highest task comprehension, as indicated by the average probability mass assigned to the labels. These label probabilities were then used as the model's confidence scores, and we examined how normalizing them (disregarding probability mass on non-label tokens) affects calibration.

Results

Evaluating the calibration plots across all configurations, we found that normalizing label probabilities is essential for meaningful confidence estimates—unnormalized scores are negatively influenced by overall task comprehension and exhibit substantially poorer calibration. Calibration error (ECE) increased with reasoning complexity: GSM8K and GPQA showed higher ECE than MMLU, indicating that token probabilities alone struggle to capture epistemic uncertainty in multi-step and symbolic reasoning tasks. Instruction tuning had a mixed impact on ECE: some models (e.g., Mistral 7B, Llama 70B) exhibited degraded calibration, while others (e.g., Mistral 24B) showed negligible change. Across all tuned models, however, we observed a pronounced polarization of confidence scores, concentrating probability mass on a single label and diminishing estimate nuance. Notably, the Ministral 8B model that was additionally assessed, although lacking a public base model, showed much lower polarization. These findings suggest that, although token-level probabilities can reliably reflect aleatoric uncertainty in straightforward tasks, they become overconfident under instruction tuning and are insufficient for quantifying uncertainty in complex, reasoning-heavy applications.

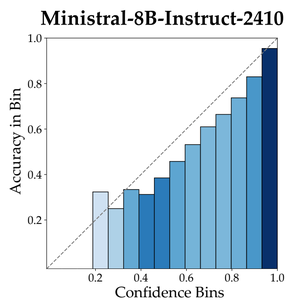
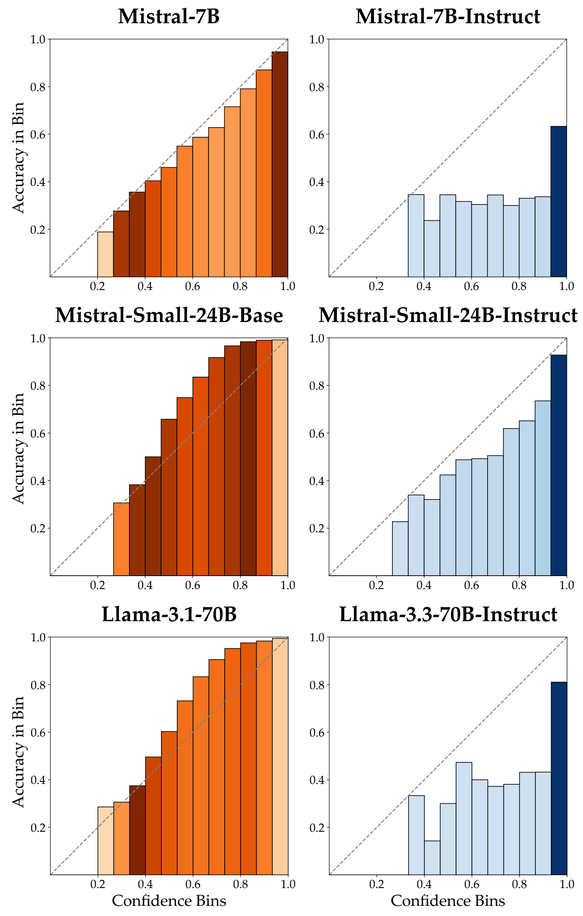


Figure 5 (left). **Calibration plot for Ministral-8B-Instruct on the same MMLU dataset.** While this model lacks a publically available base model, it shows significantly less polarization compared to the other instruction tuned models assessed.

Figure 4 (right). **Effect of Instruction Tuning on Normalized Entropy of Bin Counts.** Calibration plots for the MMLU dataset are shown for pairs of base and instruction-tuned models, using normalized label probabilities and considering only the most probable label. Base models are depicted in orange, while their instruction-tuned counterparts are shown in blue. Darker shades correspond to a higher number of items within a given bin. In instruction-tuned models, a clear shift of item counts toward the bin representing the highest confidence interval is observed. Consequently, the probabilities assigned to the other labels — which are not shown in the plot — must decrease.



Sequence Level Metric Calibration

Experiment Design

To evaluate how different uncertainty metrics perform in scientific question answering, we benchmarked four instruction-tuned models (7B, 8B, 24B, 70B) across eight datasets with clear ground truth: five multiple-choice (MMLU-Physics, ARC-Easy, ARC-Challenge, SciQ, GPQA) and three arithmetic (GSM8K, SVAMP, SciBench). Multiple-choice items were probed using counterfactual prompting with the APriCoT prompting strategy, where each choice is evaluated indiviually then classified as correct or incorrect. For arithmetic questions, Chain-of-Thought prompting was used, followed by an extraction prompt to isolate the final numeric answer. To evaluate the Frequency of Answer metrics, which requires multiple generations for the same prompt, we sampled ten generations per prompt. All datasets were subsampled to 1,000 items for tractability. This setup resulted in a total of 181,360 question answering prompts per model. We then applied four selected uncertainty metrics — Verbalized Uncertainty, P(True), Frequency of Answer, and Claim-Conditioned Probability — to compare their reliability as measured by calibration.

Summary of Results

Across the four metrics, only Frequency of Answer delivered well-calibrated confidence that closely correlated with correctness. Verbalized Uncertainty and P(True) exhibited strong response biases and minimal alignment with accuracy, whereas Claim-Conditioned Probability collapsed to near-zero scores on longer outputs due to multiplicative aggregation and NLI misclassifications. Frequency of Answer's reliability validates semantic consistency as a confidence signal, though its high sampling cost and the need for non-trivial semantic clustering make it inapplicable to open-ended QA. These findings underscore the need for more robust, efficient uncertainty estimation methods in complex, reasoning-heavy tasks.

Verbalized Uncertainty [4]

Metric Explaination: Verbalized Uncertainty prompts the model, immediately after answering, to output a numeric confidence score—typically a probability token—reflecting its self-assessed certainty.

Our results show that, although all models reliably produced valid numeric outputs, they overwhelmingly defaulted to a small set of values (e.g., 0.0, 0.5, 0.8, 0.9, 0.95, 0.99, 1.0), with higher confidences dominating the responses. This leads to a highly biased distribution of confidence score, potentially driven by training data and instruction tuning. Calibration plots reveal no meaningful correlation between these verbalized scores and answer accuracy, confirming that this method, indicating that Verbalized Uncertainty is not a reliable proxy for true model confidence.

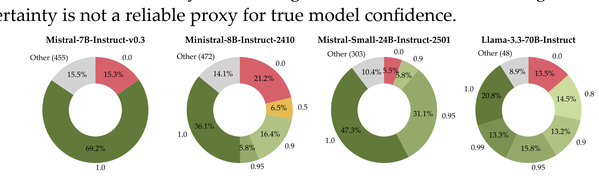
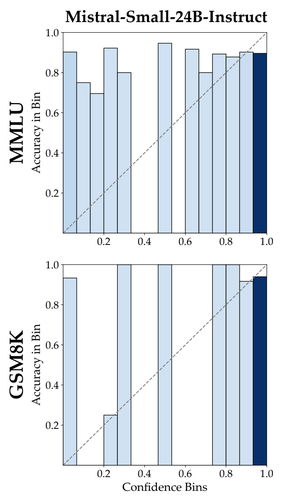


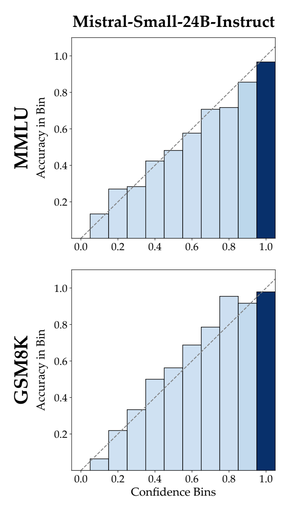
Figure 6. **Confidence Score Distribution for Verbalized Uncertainty per Model.**



Frequency of Answer

Metric Explanation: Frequency of Answer quantifies confidence by sampling multiple generations for the same prompt and assigning each answer a score equal to the frequency of semantically identical samples across all samples.

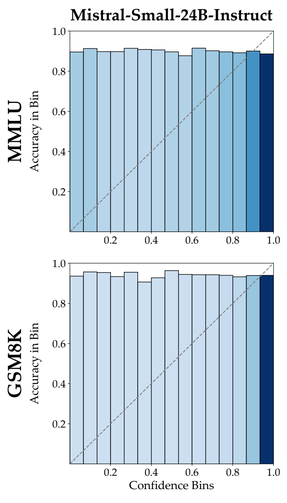
Our evaluation shows that higher answer frequencies strongly correlate with correctness across both multiple-choice and arithmetic tasks, with more challenging datasets (e.g., GPQA, SciBench) exhibiting greater answer diversity, reflecting higher model uncertainty. Calibration plots confirm well-aligned confidence estimates based on Frequency of Answer, demonstrating the reliability of this approach. However, due to its computational cost, because of requiring many samples per prompt, and its dependence on semantic clustering of outputs, it remains unapplicable to open-ended question answering.



P(True) [5]

Metric Explaination: P(True) queries the model, after producing its answer, to classify that answer as "True" or "False," using the underlying token probabilities assigned to the corresponding labels as confidence scores.

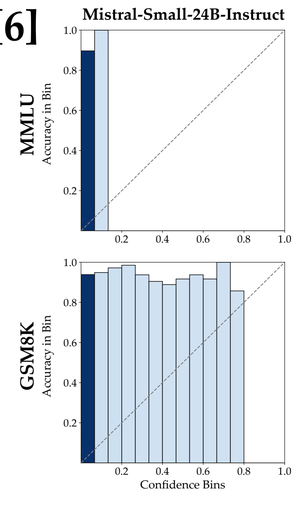
We find that P(True) suffers from pronounced response bias. In several models P(True) overwhelmingly assigns near 1.0 confidence with liŁle use of intermediate confidence scores, resulting in a polarized distribution of certainty scores. This polarization may stems from the model's commitment to a single reasoning path before classification, and varies by model. Calibration plots further reveal no meaningful correlation between P(True) scores and actual correctness, indicating that P(True) cannot reliably quantify uncertainty.



Claim-Conditioned Probability [6]

Metric Explaination: Claim-Conditioned Probability (CCP) evaluates uncertainty by, for each token in the model's output, using an NLI model to determine which of the top probable token alternatives entail the original meaning, and computing token-level certainty as the ratio of probability mass assigned to entailing tokens to the sum of entailing and contradicting tokens. Sequence-level confidence is then obtained by multiplying token certainties.

In practice, we find that CCP suffers from vanishing sequence-level scores as generation length grows, as multiplying many token certainties drives overall confidence near 0. As a result, calibration plots show no meaningful alignment with correctness. Furthermore, NLI misclassifications and the inclusion of stop words in the aggregation further destabilize scores. These issues make CCP unreliable for sequence-level uncertainty estimation, though its token-level insights could still inform targeted analyses once aggregation and domain-specific entailment are improved.



References

[1] OpenAI. GPT-4 Technical Report. 2024. arXiv: 2303.08774 [cs.CL].
[2] Ekaterina Fadeeva et al. LM-Polygraph: Uncertainty Estimation for Language Models. 2023. arXiv: 2311.07383 [cs.CL].
[3] Kyle Moore, Jesse Roberts, Thao Pham, and Douglas Fisher. Reasoning Beyond Bias: A Study on Counterfactual Prompting and Chain of Thought Reasoning. 2024. arXiv: 2408.08651 [cs.CL].
[4] Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback. 2023. arXiv: 2305.14975 [cs.CL].
[5] Saurav Kadavath et al. Language Models (Mostly) Know What They Know. 2022. arXiv: 2207.05221 [cs.CL].
[6] Ekaterina Fadeeva et al. Fact-Checking the Output of Large Language Models via Token-Level Uncertainty Quantification. 2024. arXiv: 2403.04696 [cs.CL].

Research performed at HZDR HELMHOLTZ ZENTRUM DRESDEN ROSSENDORF

with Resources generously provided by JÜLICH Forschungszentrum